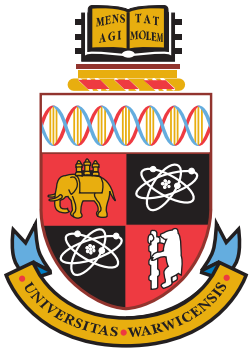


Factor regression for dimensionality reduction and data integration techniques with applications to cancer data

Alejandra Avalos Pacheco

Thesis submitted for the degree of *Doctor of Philosophy*

University of Warwick
Department of Statistics

December 2018

Contents

List of Figures

List of Tables

List of Algorithms

In loving memory of mami Meche.

Acknowledgements

Firstly I thank my supervisors, David Rossell and Richard Savage, who were always helpful, supportive, insightful and patient. I am deeply grateful to them for truly caring about my personal and professional development and well-being.

I also acknowledge David Firth and Francesco Stingo who carefully examined my thesis and provided brilliant comments and useful suggestions. They let my viva be an enjoyable moment.

I would also express my gratitude to many other Warwick, Oxford and UPF scientist and faculty members for the enlightening discussions. I particularly thank Chris Yau for valuable insights and provided data.

I am grateful to all my Warwick and Oxford friends: it has been an honour to have them in my life. I especially thank the "team Savage", who talked through ideas with me, and to Karla, Panayiota and Lewis: without them I would not have been able to submit this thesis (literally and figuratively).

Thanks to my family, in particular to my parents and papa Coco. Thanks for all the love and for always believing in me.

Thanks above all to Elia. There are no words which could express my gratitude for his support through thick and thin.

Declaration

I declare that I have written and developed this PhD thesis entitled **"Factor regression for dimensionality reduction and data integration techniques with applications to cancer data"** completely by myself, under the supervision of Dr. David Rossell and Dr. Richard S. Savage, for the degree of Doctor of Philosophy in Statistics. I have not used sources or means without declaration in the text. I also confirm that this thesis has not been submitted for a degree at any other university.

During my PhD I have written the article: **"Heterogeneous large datasets integration using Bayesian factor regression"** (Avalos-Pacheco et al., 2018), in collaboration with my supervisors Dr. David Rossell and Dr. Richard S. Savage. This article has been submitted to a peer reviewed journal and is still under revision. A preprint of this article can be found at `https://arxiv.org/abs/1810.09894`

Chapter 1 provides the necessary background for this thesis, often presented with a different notation and form in the literature (references are provided). Chapters 2, 3 and Section 4.1 are an extended version of the content of Avalos-Pacheco et al. (2018). Some of the figures of these chapters come from the aforementioned article. Section 4.2 and Chapter 5 have not been published at the time of writing.

ABSTRACT

Two key challenges in modern statistical applications are the large amount of information recorded per individual, and the fact that such data are often not collected all at once but in batches, often causing distortions in both mean and variance. We address both issues by introducing a novel sparse latent factor regression model to integrate heterogeneous data. The model provides a tool that addresses data exploration via dimensionality reduction and corrects the so-called *batch effects*, and provides sparse low-rank covariance matrix estimates. We study the use of several sparse priors, both local and non-local, to learn the dimension of the latent factors. Our model is fitted in a deterministic fashion by means of an EM algorithm for which we derive closed-form updates; this contributes a novel scalable algorithm for non-local priors, which is of interest beyond the immediate scope of this thesis. We also present several examples, with a focus on bioinformatics applications. Our results mainly show an increase in the accuracy of low-dimensional data reconstructions, with non-local priors substantially improving the inference on factor cardinality and non-zero factor loadings. Moreover, thanks to our batch effect correction, we achieve a considerable improvement in recovering the latent factors. Altogether, this thesis provides a novel approach to latent factor regression that balances sparsity with sensitivity, as well as being highly computationally efficient, and opens new avenues for future research on dimension-reduction-based data integration. The methodology developed in this thesis is available in an **R** package at `https://github.com/AleAviP/BFR.BE`.

1

Introduction

1.1 Motivation

New technologies enable the gathering of large datasets. While these offer great promise for science, policy making and industry, their large volume and in particular the large number of recorded variables make its analysis and interpretation more challenging.

Two main challenges in dealing with this high volume of data from a statistical perspective are:

(i) The *high dimensional nature of the data* often leads to models with a large number of parameters. These can be hard to handle, may obstruct interpretability, and often require computationally intensive calculations. A first important task when studying large datasets is to conduct an exploratory analysis: dimensionality reduction techniques have proven to be a highly popular tool for this purpose.

(ii) *Batch effects*, i.e. systematic biases in data that are unrelated to the scientific signal of interest, might arise when data are generated under different experimental conditions, when new samples are incrementally added to existing samples, or in analyses coming from different projects, laboratories, or platforms. Batch effects may lead to biased or inaccurate inference, unless properly adjusted for (Leek et al., 2010; Goh et al., 2017).

We address these challenges via a framework for integrative models based on Bayesian factor analysis and latent factor regression, resulting into sparse latent factors. We also provide a mean and variance batch effect correction via a location-scale adjustment. We thus develop a model-based approach (the first, as far as we know) for tackling dimensionality reduction and batch effect correction simultaneously. The main contributions of our work are:

(i) A novel and scalable non-local prior based formulation to induce sparsity and learn the underlying number of factors, including important aspects related to prior elicitation. To our knowledge this is the first adaptation of non-local priors to factor models.

(ii) A flexible model-based Bayesian factor regression approach for correcting batch effects.

(iii) An efficient and scalable Expectation Maximisation algorithm with closed-form updates to obtain the Maximum A Posteriori (MAP) parameter estimates.

(iv) An R implementation publicly available at `https://github.com/AleAviP/BFR.BE`.

(v) Application of these new methods to a variety of real-world and synthetic data sets.

As we will see, non-local priors provide a good balance between sparsity and sensitivity in inferring non-zero loadings; moreover, they give a better estimation of factor cardinality than other similar sparse inducing priors, even in scenarios without batch effects.

In this thesis we will focus on cancer-related gene expression data as our main motivating application. Cancer is one of the most studied pathological systems and one of the leading causes of morbidity and mortality worldwide. In 2012 about 14.1 million new cases occurred globally and about 8.2 million people died from cancer, corresponding to 14.6% of all human deaths (World Health Organization, 2014). It is expected that the number of annual cases will rise from 14 million in 2012 to 22 million within the next 20 years. Cancer cases occur more commonly in developed countries and risk increases significantly with age. In the UK, for example, more than one in three people will develop some form of cancer during their lifetime (NHS, 2015); the survival rate is around 50%, and it is estimated that 42% of the cases could be prevented (Cancer Research UK, 2015).

Dimensionality reduction techniques are a popular tool for a better understanding of cancer (Gligorijevic and Przulj, 2015) and can lead to a more comprehensive analysis and easier interpretation of the data (Kristensen et al., 2014). On the other hand, batch effects are a major issue that arises in cancer datasets (Choi et al., 2017) and needs to be corrected in order to obtain accurate conclusions (Johnson and Li, 2009; Zhu et al., 2017).

Let us also notice that batch effects are present in many other research fields, such as structural magnetic resonance imaging (MRI) data from Alzheimer's disease (Shinohara et al., 2014; Fortin et al., 2016), multiple sclerosis (Shah et al., 2011), and attention deficit hyperactivity disorder (Olivetti et al., 2012), or even toxicological studies in marine species (Avio et al., 2015).

This chapter starts by reviewing dimensionality reduction, sparsity, non-local priors,

and batch effect correction techniques: we present the current state-of-the-art and provide the necessary background for our model.

1.2 Dimensionality Reduction

Dimensionality reduction techniques represent the data into a low-dimensional Euclidean space that gives insight into its underlying structure in order to visualise, denoise or extract meaningful features.

Let $X \in \mathbb{R}^{n \times p}$ be a data matrix, where entry x_{ij} is the i^{th} observation corresponding to the j^{th} variable ($i = 1, \ldots, n$ and $j = 1, \ldots, p$) and x_i^T denotes the i^{th} row. We will refer to X as the high-dimensional matrix, alluding to the fact that the number of variables p is potentially large. Without loss of generality, it will be assumed from now on that X has zero column means. Let $Z \in \mathbb{R}^{n \times q}$ be a low-dimensional matrix (in the sense that $q \ll p$), with z_i^T denoting the i^{th} row.

The goal of dimensionality reduction methods is, given $X \in \mathbb{R}^{n \times p}$, to obtain $Z \in \mathbb{R}^{n \times q}$ ($q \ll p$) that in some respect provides a useful and more compact representation of the original data. This is done by finding a function $Z = f(X)$, requiring that the low-dimensional representation possesses some desirable properties, such as preserving variance or capturing covariance. Typically, given a pre-specified class of functions $\mathcal{F}$, one searches for a function f that achieves

$$\min_{f \in \mathcal{F}} g(X, f(X)) \tag{1.1}$$

where $g(\cdot)$ measures the quality of the low-dimensional representation. We remark that g might be some distance or discrepancy measure, as in the classical optimisation-based approaches to dimensionality reduction, but can also arise from a likelihood or posterior density function in model-based frameworks.

Principal component analysis and factor analysis are two classical dimensionality reduction methods. Factor analysis is an extension of principal component analysis, thus in this thesis we focus on the latter for the sake of generality. In particular, we study Bayesian factor analysis and we address two key challenges: sparsity and scalable computation. For other dimensionality reduction methods, see Johnson and Wichern (1988, chap. 3) or Hastie et al. (2001, chap. 14) for a gentle introduction, and Burges (2010); Cunningham and Ghahramani (2015) for more recent reviews.

1.2.1 Principal Component Analysis

Principal Component Analysis (PCA) is a linear dimensionality reduction technique developed by Pearson (1901) to transform the high-dimensional data X into a low-dimensional representation Z, such that Z is a linear uncorrelated transformation of X that conserves as much variance as possible.

PCA minimises the reconstruction error via least squares optimisation:

$$\begin{aligned} &\min ||X - ZM^\top||^2 \\ &\text{subject to } M^\top M = I, \end{aligned} \tag{1.2}$$

with $M \in \mathbb{R}^{p \times q}, q \leq p$.

Equation (1.2) is optimised when

$$\widehat{Z} = X\widehat{M}(\widehat{M}^\top \widehat{M})^{-1} = X\widehat{M}. \tag{1.3}$$

We are left to find the matrix M. The optimal solution to (1.2) is obtained via the Singular Value Decomposition (SVD) of X:

$$X = ULV^\top \tag{1.4}$$

where U, called left singular vectors, is an $n \times n$ orthogonal matrix ($U^\top U = I$); V, called right singular vectors, is a $p \times p$ orthogonal matrix ($V^\top V = I$) and L is an $n \times p$ diagonal matrix with non-negative real diagonal entries called singular values $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_{\min(n,p)} \geq 0$. In terms of this factorisation, the covariance matrix

$$\Sigma = X^\top X = VL^2V^\top \tag{1.5}$$

has the same right singular vectors V as X. Using the SVD, we have that $Z = UL^{1/2}$ and $M = L^{1/2}V^\top$, and

$$\widehat{M} = V_q L_q^{1/2} \tag{1.6}$$

gives the rank q solution, where A_q are the first q columns of a given matrix A, V_q contains the q eigenvectors (called singular vectors) corresponding to the largest eigenvalues, and $\widehat{Z} = X\widehat{M}$ are known as principal components. Figure 1 provides a visual representation of PCA, with $p = 2$ and $q = 1$.

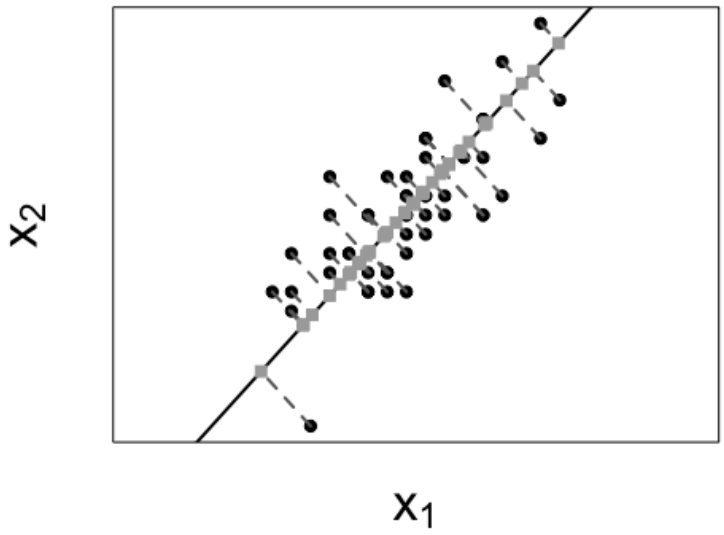


Figure 1. Principal component analysis representation: 2-dimensional variables X in black points and 1-dimensional representations Z in squared grey points.

1.2.2 Probabilistic Principal Component Analysis

Probabilistic Principal Component Analysis (PPCA) (Tipping and Bishop, 1999) adds a probabilistic generative model to PCA. Observations X are seen as linear combinations of principal components Z plus an $n \times p$ matrix of errors E, where $\mathsf{e}_i^\top$ denotes the i^{th} row and $e_{ij} \sim N(0, \sigma_\varepsilon^2)$ are independent across $i = 1, \ldots, n, j = 1, \ldots, p$:

$$\mathsf{x}_i = M\mathsf{z}_i + \mathsf{e}_i. \tag{1.7}$$

Note that PPCA assumes the variance to be constant across (i, j): this assumption is relaxed by factor models by allowing sigma to depend on j.

The low-dimensional representations are independent standard normal random variables $\mathsf{z}_i \sim N(0, \mathbf{I}_q)$, for $i = 1, \ldots, n$.

Thus, $\mathsf{p}(\mathsf{x}_i \mid \mathsf{z}_i, M, \sigma_\varepsilon) = N(Mz_i, \sigma_\varepsilon^2 \mathbf{I}_q)$. Integrating out z_i we obtain the marginal

model $\mathsf{p}(\mathsf{x}_i \mid M, \sigma_\varepsilon) = N(0, MM^\top + \sigma_\varepsilon^2 \mathbf{I}_p)$, leading to the log-likelihood:

$$\log \mathsf{p}(\mathsf{x}_i \mid M, \sigma_\varepsilon) \propto -\frac{1}{2} \sum_{i=1}^{n} \left[\mathsf{x}_i^\top (MM^\top + \sigma_\varepsilon^2 \mathbf{I})^{-1} \mathsf{x}_i \right] - \frac{n}{2} \log |MM^\top + \sigma_\varepsilon^2 \mathbf{I}|. \tag{1.8}$$

Notably Equation (1.8) can be maximised in closed-form. Specifically, maximisation with respect to M gives

$$\widehat{M} = V_q (L_q - \sigma_\varepsilon^2 \mathbf{I})^{1/2}, \tag{1.9}$$

where $\widehat{\Sigma} = \frac{1}{n} X^\top X = V^\top L^2 V$ is the SVD, and the maximum likelihood estimator for σ_ε^2 is

$$\widehat{\sigma}_\varepsilon^2 = \frac{1}{(p-q)} \sum_{j=q+1}^{p} \lambda_j. \tag{1.10}$$

We note that $\mathsf{p}(\mathsf{z}_i \mid \mathsf{x}_i, M, \sigma_\varepsilon) = N((M^\top M + \sigma_\varepsilon^2 \mathbf{I}_q)^{-1} M^\top \mathsf{x}_i, \sigma_\varepsilon^2 (M^\top M + \sigma_\varepsilon^2 \mathbf{I}_q)^{-1})$. Thus,

$$\begin{aligned} \mathbb{E}[\mathsf{z}_i \mid \mathsf{x}_i, \widehat{M}, \widehat{\sigma}_\varepsilon] &= (\widehat{M}^\top \widehat{M} + \widehat{\sigma}_\varepsilon^2 \mathbf{I}_q)^{-1} \widehat{M}^\top \mathsf{x}_i \\ &= (\mathbf{I}_q + \widehat{\sigma}_\varepsilon^2 \mathbf{I}_q)^{-1} \widehat{M}^\top \mathsf{x}_i, \end{aligned} \tag{1.11}$$

since $M^T M = \mathbf{I}$.

Then, PPCA can be seen as a ridge-type regression of z_i versus x_i with penalty σ_ε^2, recovering PCA when σ_ε^2 is zero.

Extensions of this method include Ulfarsson and Solo (2008), who added a Gaussian shrinkage prior $p(M)$ on the entries of M, and Kao and Roy (2013), who imposed shrinkage by a penalty on the inverse covariance matrix.

1.2.3 Factor Analysis

Factor analysis is a dimensionality reduction technique that aims to describe the covariance of an observed set of variables. It was originally introduced by Spearman (1904) as a tool for psychometric analysis. The observed variables $\mathsf{x}_i \in \mathbb{R}^p, i = 1, \ldots, n$, are modelled as in (1.7), but here e_i are independent Gaussian $\mathsf{e}_i \sim N(0, \mathcal{T}^{-1})$, where $\mathcal{T}$ is a $p \times p$ diagonal matrix, z_i and e_i are independent, $\mathsf{z}_i \sim N(0, \mathbf{I}_q)$. In this context variables, z_i are called latent factors, $M \in \mathbb{R}^{p \times q}$ is the matrix of factor loadings, and $\mathcal{T}^{-1}$ are the idiosyncratic variances (Fruchter, 1955).

Thus $\mathsf{x}_i \mid \mathsf{z}_i, M, \mathcal{T} \sim N(M\mathsf{z}_i, \mathcal{T}^{-1})$, for a diagonal matrix $\mathcal{T}$ and the marginal distribu-

tion of x_i is $\mathsf{x}_i \mid M, \mathcal{T} \sim N(0, MM^\top + \mathcal{T}^{-1})$. Hence, the log-likelihood is:

$$\log \mathrm{p}(X \mid M, \mathcal{T}) \propto -\frac{1}{2} \sum_{i=1}^{n} \left[\mathsf{x}_i^\top (MM^\top + \mathcal{T}^{-1})^{-1} \mathsf{x}_i \right] - \frac{n}{2} \log |MM^\top + \mathcal{T}^{-1}|. \tag{1.12}$$

The conditional distribution of the latent factors is

$$\mathsf{z}_i \mid \mathsf{x}_i, M, \mathcal{T} \sim N((M^\top \mathcal{T}^{-1} M + \mathbf{I}_q)^{-1} M^\top \mathcal{T} \mathsf{x}_i, (M^\top \mathcal{T}^{-1} M + \mathbf{I}_q)^{-1}). \tag{1.13}$$

Thus, conditional on some estimated $(\widehat{M}, \widehat{\mathcal{T}})$, a simple option to obtain low-dimensional coordinates is to use

$$\mathbb{E}[\mathsf{z}_i \mid \mathsf{x}_i, \widehat{M}, \widehat{\mathcal{T}}] = (\widehat{M}^\top \widehat{\mathcal{T}}^{-1} \widehat{M} + \mathbf{I}_q)^{-1} \widehat{M}^\top \widehat{\mathcal{T}} \mathsf{x}_i. \tag{1.14}$$

However, obtaining $(\widehat{M}, \widehat{\mathcal{T}})$ via MLE or posterior mode estimation cannot be done in closed-form as in PPCA, hence a numerical optimisation scheme such as the Expectation-Maximisation algorithm is used to estimate them.

Notice that, when the observation noise is $\mathcal{T}^{-1} = \sigma_\varepsilon^2 I$, we recover PPCA, whereas for $\mathcal{T}^{-1} = 0$ we obtain PCA. Thus, we will focus on FA for the sake of generality.

Many extensions of Factor Analysis have been developed. For instance, Geweke and Zhou (1996) developed a Bayesian model for factor analysis, providing a Gibbs sampler for inference. Lopes and West (2004) provided a Bayesian framework where the factor model and the number of factors q are determined simultaneously; here the inference is made via reversible jump MCMC algorithms. Carvalho et al. (2008) provided a sparse FA and factor regression model which is applied to breast cancer: data are decomposed into latent factors – representing biological "subpathway"– on one side, and known biological information and clinical biomarkers on the other side, using non-Gaussian/non-parametric sparsity-inducing priors. Other recent works on Bayesian exploratory FA are: Kao and Roy (2013), who introduced a probabilistic FA method; Hirose and Yamamoto (2015), who proposed sparse factor models with a nonconcave penalised likelihood, and Ročková and George (2017) who provided a model with automatic rotations to sparsity for the loadings, with a parameter-expanded Expectation-Maximisation algorithm for inference. In Chapter 2 we extend the model of Ročková and George (2017), increasing the flexibility of their work to account for batch effects, obtaining sparse factors via non-local priors. Such an extension addresses two main challenges: sparsity and scalable computation.

1.3 Sparsity

From a statistical point of view it is often desirable to enforce sparse solutions to improve interpretability and, when the number of parameters to be estimated is large, potentially also improve the accuracy of statistical inference. Moreover we argue that, as an additional motivation, sparsity is a bona-fide prior expectation in certain applications. For instance, in genetics a few active genes could potentially explain a whole complex biological system; e.g. it was found that few latent variables associated to the cerebrum tissue of mice could be used to provide their genomic "true age" (Perry and Owen, 2010). In psychology, some social behaviours could be explained via latent factors: four of them were found to explain 81% of the total variance in a job scoring (Kendall, 1975). In social media, for example, the most popular videos played in YouTube were produced by 10,000 out of 1 billion users (Earnshaw, 2017).

Several strategies have been developed for sparse models, such as Least Absolute Shrinkage and Selection Operator (LASSO) regression (Tibshirani, 1994), Strawderman - Berger priors (Strawderman, 1971; Berger, 1980), nonconcave penalties e.g. smoothly clipped absolute deviation (SCAD) penalty (Fan and Li, 2001) or minimax concave penalty (MCP) (Zhang, 2010), and Horseshoe priors (Carvalho et al., 2009). Another approach is to model the loadings via spike-and-slab priors (SS), which are a mixture of two distributions: one for the important loadings (slab) and another one for the non-important loadings (spike). One possibility is to model the spike component with a finite point mass on zero and the slab with a continuous density (Mitchell and Beauchamp, 1988; West, 2003; Knowles and Ghahramani, 2011).

In this thesis we focus on continuous SS models, estimating them via Expectation-Maximisation (EM) algorithms, which induce posterior zeroes on the loadings with high probability. As a starting point, we take the EM algorithm developed in Ročková and George (2017). The continuity of the spike distribution enables us to obtain closed-form expressions for the EM updates, providing a rapidly computable and scalable EM algorithm.

Let $\gamma_{jk} \in \{0, 1\}$ be latent indicators for $j = 1, \ldots, p$ and $k = 1, \ldots, q$, with $\gamma_{jk} = 1$ if latent factor z_{jk} is included in the model. In the SS models, the inclusion of the latent factors z_{jk} is modelled through the loadings m_{jk}. Thus, when $m_{jk} = 0$, we set $z_{jk} = 0$, distinguishing the latent factors to be excluded from the ones to be included. The loadings are modelled as follows:

$$\mathrm{p}(m_{jk} \mid \gamma, \lambda_0, \lambda_1) = (1 - \gamma_{jk})\, \mathrm{p}(m_{jk} \mid \lambda_0, \gamma_{jk} = 0) + \gamma_{jk}\, \mathrm{p}(m_{jk} \mid \lambda_1, \gamma_{jk} = 1), \tag{1.15}$$

where λ_0 and λ_1 are the dispersion parameters of the spike and slab components respectively, with $\lambda_1 > \lambda_0$.

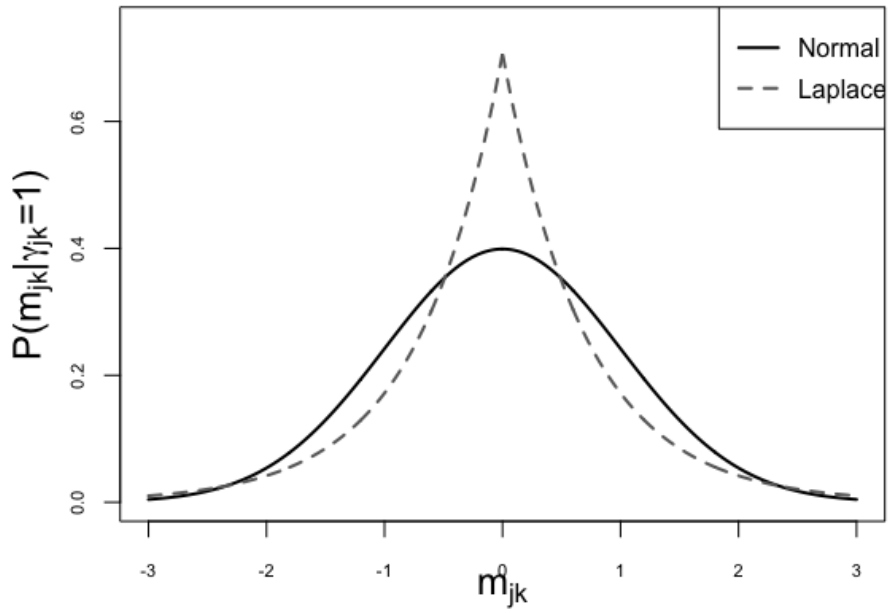


Figure 2. Local priors for m_{jk} under a model $\gamma_{jk} = 1$.

Two popular densities in SS models for modelling the non-zero loadings are Normal (George and McCulloch, 1993) and Laplace (Ročková and George, 2018). Figure 2 shows the slab priors used to model the non-zero loadings therein. Notice that the regions of the real line to which the spike and the slab assign positive non-negligible probability "overlap" in a neighbourhood around $m_{jk} = 0$: thus, as we will see later (Figures 8 and 10 right panels), it becomes hard to tell which of them (spike or slab) generated m_{jk}. To solve these issues, we then model the non-zero loadings with non-local priors.

1.4 Non-local priors

Non-local priors (NLPs) were introduced by Johnson and Rossell (2010) and aim to enforce separation between competing probability models. From a practical point of view, NLPs lead to stronger parsimony and, as we illustrate later, in certain situations can also help increase sensitivity to detect non-zero coefficients. NLPs are appealing in our FA

setting because, when modelling the non-zero loadings, they vanish as m_{jk} gets close to zero. In this context, any other prior that does not vanish as $m_{jk} \to 0$ is called Local Prior (LP).

We would like to model the probability $\mathrm{p}(m_{jk} \mid \gamma_{jk} = 0)$ for non-important loadings ($m_{jk} = 0$) and the probability $\mathrm{p}(m_{jk} \mid \gamma_{jk} = 1)$ for important loadings ($m_{jk} \neq 0$).

An NLP density can always be expressed as

$$\mathrm{p}(m_{jk} \mid \gamma_{jk} = 1) = d(m_{jk})\, \mathrm{p}_L(m_{jk} \mid \gamma_{jk} = 1), \tag{1.16}$$

where $d(m_{jk})$ is a penalty term and $\mathrm{p}_L(m_{jk} \mid \gamma_{jk} = 1)$ is a base local prior density (Rossell and Telesca, 2017).

Some default additive NLPs are:

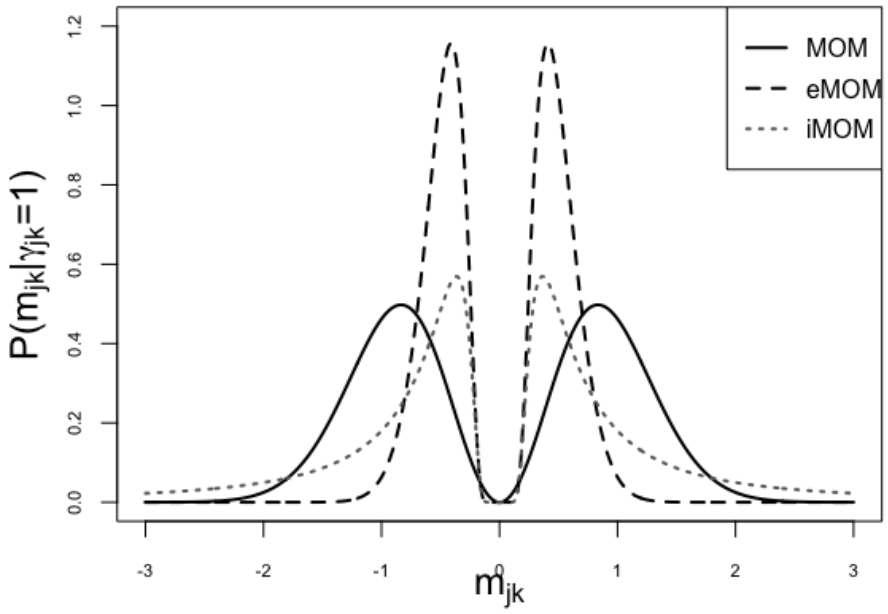


Figure 3. Non-local priors for m_{jk} under a model $\gamma_{jk} \neq 0$.

(i) **Moment (MOM)** (Johnson and Rossell, 2010). These priors are obtained as the product between the (2κ)-th power of the parameter of interest m_{jk} and a base local prior

density $p_L(m_{jk})$ with 2κ finite integer moments. The MOM density is then

$$p_M(m_{jk}) = \frac{m_{jk}^{2\kappa}}{\Upsilon} p_L(m_{jk}), \tag{1.17}$$

where

$$\Upsilon = \int m_{jk}^{2\kappa} p_L(m_{jk}) dm_{jk}$$

is called prior dispersion parameter. Under some conditions, MOM priors lead to closed-form expressions for MCMC algorithms and, as we will see in Chapter 2, for EM algorithms.

In this thesis, we consider the base local prior densities $p_L(m_{jk})$ to be either Normal or Laplace. The resulting MOM densities are the Normal-based prior:

$$\frac{m_{jk}^2}{\widetilde{\lambda}_1} N(m_{jk}; 0, \widetilde{\lambda}_1), \tag{1.18}$$

with $\Upsilon = \widetilde{\lambda}_1$, and the Laplace-based prior:

$$\frac{m_{jk}^2}{2\widetilde{\lambda}_1^2} \text{Laplace}(m_{jk}; 0, \widetilde{\lambda}_1), \tag{1.19}$$

with $\Upsilon = 2\widetilde{\lambda}_1^2$.

(ii) **Inverse moment (iMOM)** (Johnson and Rossell, 2010). iMOM densities are of the form

$$p_I(m_{jk}) = \frac{\kappa \widetilde{\lambda}_1^{\varpi/2}}{\Gamma(\varpi/2\kappa)} \left(m_{jk}^2\right)^{-(\varpi+1)/2} \exp\left[-\left(\frac{m_{jk}^2}{\widetilde{\lambda}_1}\right)^{-\kappa}\right], \tag{1.20}$$

for $\varpi, \widetilde{\lambda}_1 > 0$. An iMOM prior has a similar form as an Inverse Gamma when $m_{jk} \to 0$. As we can see in Figure 3, the iMOM prior approaches zero faster than the MOM prior. The drawback of iMOM priors is that they do not lead to closed-form expressions for MCMC and EM algorithms.

(iii) **Exponential Moment (eMOM)** (Rossell et al., 2013). eMOM priors are

$$p_E(m_{jk} \mid \gamma_{jk} = 1) = e^{\sqrt{2}} \exp\left[-\frac{\widetilde{\lambda}_1}{m_{jk}^{2\kappa}}\right] N(m_{jk}; 0, \widetilde{\lambda}_1). \tag{1.21}$$

Unlike iMOM priors, eMOM priors provide closed-form expressions under certain MCMC setups, similarly to MOM priors. The main difference between MOM and

eMOM priors is that the former vanish at a polynomial speed as $m_{jk} \to 0$, whereas the latter vanish exponentially fast. In our experience MOM priors lead to simpler computation and, provided the prior parameters are suitably elicited, are enough to induce sufficient sparsity.

As a brief review, NLPs have been applied to Bayesian model selection (BMS) and model averaging (BMA) in linear regression (Johnson and Rossell, 2010, 2012; Rossell and Telesca, 2017); in some generalised linear models (Johnson and Rossell, 2012; Rossell et al., 2013); under orthogonal and block-diagonal regression (Papaspiliopoulos and Rossell, 2017); for mixture models (Fúquene et al., 2018); in linear regression with non-normal residuals (Rossell and Rubio, 2018). Finally, see Shi et al. (2018) for a more recent work in parallel that also applies Spike-and-slab priors with a MOM slab component for BMS in linear regression via Gibbs sampling. Outside the generalised linear model framework, NLPs have been also studied for directed acyclic graphs (Consonni and La Rocca, 2011), gene regulatory networks (Chekouo et al., 2015), chain event graphs (Collazo and Smith, 2016), and Bayesian graphical regression (Ni et al., 2018). In Chapter 2 we study Normal-based MOM priors in the FA setting, but we also discuss some Laplace-tailed extensions. To our knowledge this is the first adaptation of NLPs to factor models. As we will discuss later, the main advantage of NLPs in this setting is to help achieve a better balance between sparsity and sensitivity in inferring non-zero loadings.

1.5 Combining data and batch effects

When data from multiple sources, projects or experiments are available, one would like to perform a statistical analysis that incorporates all available information. In this setting it is important to take into account batch effects, i.e. systematic artefacts that may lead to imprecise or erroneous findings. Several batch effect correction algorithms (BECA) have been developed (see Scherer (2009); Lazar et al. (2013) for a review and examples). They can be divided into data "normalisation", matrix factorisation-based and location-scale methods. We overview some popular strategies and discuss their advantages and limitations.

We highlight that all the techniques described in this section are pre-processing methods only. Any downstream analysis performed *after* obtaining the batch effect adjusted data does not take into account that the data have been pre-processed. Therefore, the uncertainty or any possible errors in the pre-processing are ignored. This thesis, instead, proposes a model that carries out dimension reduction and batch effect correction simultaneously.

1.5.1 Data "normalisation"

The main idea behind data "normalisation" is to use control metrics or regression methods to correct for the high variability due to systematic noise and artefacts.

Linear normalisation is one of the most in demand algorithm, among data normalisation BECA. Its main assumption is that the different observations $\mathrm{x}_i \in \mathbb{R}^p$ are related linearly with a baseline experiment $\widetilde{\mathrm{x}} \in \mathbb{R}^p$, as a straight line with a zero y-intercept. The method simply linear regresses $\widetilde{\mathrm{x}}$ on x_i, obtaining a scaling factor β (slope of the line):

$$\widetilde{\mathrm{x}} \approx \beta \mathrm{x}_i. \tag{1.22}$$

Other data normalisation algorithms include non-linear normalisation. See Schadt et al. (2001); Yang et al. (2002) for a review of data normalisation BECA.

"Normalisation" methods are useful but could fail when the variation between batches is large (Johnson and Li, 2009).

1.5.2 Matrix factorization-based methods

Matrix factorization-based BECA are grounded on the assumption that the most important source of variability is associated with batches. Batch effect correction is performed in two steps:

(i) concatenate the data and perform a matrix factorization;

(ii) remove the factors associated with the batches and reconstruct the data.

Alter et al. (2000); Leek and Storey (2007) proposed an adjustment using singular value decomposition, and Benito et al. (2004) via distance weighted discrimination.

These approaches present some disadvantages. It might not be straightforward to identify the batch effect component, which can be confounded with other sources of variance. Furthermore, these methods could potentially fail in cases where batches have small sample sizes or in the presence of many more than three batches (Johnson and Li, 2009).

1.5.3 Location-scale methods

Location-scale (LS) methods aim to standardise the data so that each batch displays the same or similar mean and/or variance per gene. We outline some of the most popular LS tools.

Let x_{ij} be the observed data of individual $i = 1, \ldots, n$ for variable $j = 1, \ldots, p$. Assume there are n_l individuals in batch l, so that $n = n_1 + \cdots + n_{p_b}$. Let b_i be the indicator vector of length p_b defined as $b_{il} := 1$ if individual i is in batch l, $b_{il} := 0$ otherwise.

Batch-mean centring

This method performs a mean batch correction only (Sims et al., 2008). It obtains a new corrected observation $\widehat{x}_{ij}$ such that

$$\widehat{x}_{ij} = x_{ij} - \bar{x}_{jl}, \tag{1.23}$$

where $\bar{x}_{jl} = \frac{1}{n_l} \sum_{i=1}^{n} x_{ij} b_{il}$ is the sample mean of variable j in batch l.

Ratio-based methods

These methods are a straightforward extension of batch-mean centring. They subtract the geometric mean per batch, which is less sensitive to outliers (Novoradovskaya et al., 2004), instead of the sample mean:

$$\widehat{x}_{ij} = x_{ij} - \sqrt[n_l]{\prod_{i=1}^{n} x_{ij} b_{il}}. \tag{1.24}$$

The median or arithmetic mean ratio can be also used instead.

Data standardisation

Li and Wong (2001) consider that batches may affect not only means but also variances. They therefore propose normalising the data to zero mean and unit variance per batch :

$$\widehat{x}_{ij} = \frac{x_{ij} - \bar{x}_{jl}}{\widehat{\sigma}_{jl}}, \tag{1.25}$$

where $\widehat{\sigma}^2_{jl} = \frac{1}{n_l} \sum_{i=1}^{n} (x_{ij} b_{il} - \bar{x}_{jl})^2$ is the estimated variance of variable j in batch l.

Regression-based LS adjustments

These BECA aim to mean center and standardise the variance for each variable per batch independently via a linear regression model:

$$x_{ij} = \alpha_j + \beta_j^\top b_i + \delta_j^\top b_i \varepsilon_{ij}, \tag{1.26}$$

where α_j is the overall intercept, $\beta_j \in \mathbb{R}^{p_b}$ and $\delta_j \in \mathbb{R}^{p_b}$ are the additive and the multiplicative batch effect adjustment for variable j respectively, and $\varepsilon_{ij} \sim N(0, \sigma_j^2)$ are the errors.

Some recent LS methods are the ones proposed by Leek and Storey (2007); Parker et al. (2014); Hornung et al. (2016). A more general approach, developed by Li and Wong (2003), consists in incorporating covariates $\mathsf{v}_i \in \mathbb{R}^{p_\upsilon}$ of interest. Here data are modelled as

$$x_{ij} = \alpha_j + \theta_j^\top \mathsf{v}_i + \beta_j^\top b_i + \delta_j^\top b_i \varepsilon_{ij}, \tag{1.27}$$

with $\theta_j \in \mathbb{R}^{p_\upsilon}$ the regression coefficients. These covariates help to incorporate useful information about the data (e.g. standard medical history) and to combine data from more diverse data sources (e.g. different platforms). The corrected data are

$$\widehat{x}_{ij} = \frac{x_{ij} - \widehat{\alpha}_j - \widehat{\theta}_j^\top \mathsf{v}_i - \widehat{\beta}_j^\top b_i}{\widehat{\delta}_j^\top b_i} + \widehat{\alpha}_j + \widehat{\theta}_j^\top \mathsf{v}_i, \tag{1.28}$$

with $\widehat{\alpha}, \widehat{\beta}, \widehat{\theta}, \widehat{\delta}$ the estimators for $\alpha, \beta, \theta, \delta$.

Empirical Bayes: ComBat

The previous BECA are useful, but require that the sample size is large enough so that (α, θ, β) and particularly the variance batch effect δ^2 are estimated precisely. Johnson and Li (2009) proposed a method, called empirical Bayes batch effect correction (ComBat), which considers the model in (1.27), correcting the data via three steps:

(i) **Data standardisation**. Mean and variance variable-wise data standardisation of the form:

$$y_{ij} = \frac{x_{ij} - \widehat{\alpha}_j - \widehat{\theta}_j^\top \mathsf{v}_i}{\widehat{\sigma}_j}, \tag{1.29}$$

using the least squares estimates.

(ii) **Empirical Bayes parameter estimation**. The standardised variables are assumed to be distributed as $y_{ij} \sim N(\beta_{jl}, \delta_{jl}^2)$, for $b_{jl} = 1$. The prior specification is completed with $\beta_{jl} \sim N(\mu_j, \tau_j^2)$ and $\delta_{jl}^2 \sim$ Inverse Gamma(λ_j, β_j). Such priors were selected due to their conjugacy. Hyperparameters $\mu_j, \tau_j^2, \lambda_j, \beta_j$ are estimated empirically, and estimators for β_{jl} and δ_{jl}^2 are obtained by their conditional posterior means.

(iii) **Data correction**. Data are corrected using the estimations obtained in Step (ii):

$$\widehat{x}_{ij} = \frac{\widehat{\sigma}_j}{\widehat{\delta}_j^\top b_i}(y_{ij} - \widehat{\beta}_j^\top b_i) + \widehat{\alpha}_j + \widehat{\theta}_j^\top \mathsf{v}_i. \tag{1.30}$$

ComBat is a popular tool for batch effect correction, as it is robust to outliers in small

sample sizes, improves precision, and avoids over-correcting the data. These are due to the fact that it uses empirical Bayes estimates and a hierarchical model, which borrows strength in the estimation of δ across batches. Due to these advantages, we will compare our models with ComBat throughout this thesis.

Outline of the thesis. Chapter 2 reviews the Bayesian factor analysis model and introduces our novel non-local based priors on the loadings to induce sparsity. Chapter 3 extends the Bayesian FA developed in Chapter 2 to our general framework: a Bayesian factor regression model which includes mean and variance batch effect adjustment. Chapters 2 and 3 also explain important aspects related to prior parameter elicitation; describe several EM algorithms for model fitting; provide parameter initialisation and post-processing steps required for effective model selection and dimension reduction; and set up simulations to assess the accuracy of our models. Chapter 4 presents applications on public and previously unpublished cancer datasets, under unsupervised and supervised settings. Chapter 5 presents possible extensions for future research, outlining more flexible models that assume batch specific factors or other types of data integration. Chapter 6 concludes.

2

Bayesian factor analysis with a novel Spike-and-Slab prior

2.1 Introduction

In this chapter we study dimensionality reduction via Bayesian factor analysis. Factor analysis had proven to be an efficient tool to obtain low-dimensional latent representations of the high-dimensional data by extracting latent variables (or factors) from the data. Such factors aim to provide a better understanding of the complex data, generating visual representations, new meaningful features extracted from the data or denoised latent variables.

We present a novel type of continuous spike-and-non-local-slab prior to estimate the latent cardinality (number of factors). Those priors are based on Johnson and Rossell (2010, 2012) and on the continuous Gaussian spike-and-slab prior of George and McCulloch (1997); Ročková and George (2014) and its Laplace-based extention (Ročková and George, 2017, 2018).

We provide an Expectation-Maximisation (EM) algorithm to obtain the Maximum A posteriori (MAP) estimation of the parameters. We provide closed-form EM updates, giving a novel scalable algorithm for non-local priors. To our knowledge this is the first time non-local priors are implemented in factor analysis settings.

As we will discuss later, the main advantage of non-local priors in this setting is to help achieve a better balance between sparsity and sensitivity in inferring non-zero loadings. See also Bar et al. (2018) who argued for improved sensitivity via 3-component mixture priors that resemble non-local priors in generalised linear models.

This chapter is organised as follows: Section 2.2 provides an overview to Bayesian factor analysis. Section 2.3 discuss the choice of the cardinality of the factors q. Section 2.4 reviews the continuous spike-and-slab priors. Section 2.5 gives an EM algorithm

for Normal-spike-and-slab priors and Section 2.6 for Laplace-spike-and-slab. Section 2.7 introduces non-local priors in the factor analysis context. Sections 2.8 and 2.9 provide our novel EM algorithms for our novel Normal-spike-and-MOM-slab and Laplace-spike-and-MOM-slab respectively. Section 2.10 provides guidelines for the prior elicitation. Section 2.11 and 2.12 discuss the initialisation and post-processing of the parameters. Section 2.13 shows the potential of our model with simulated data. Finally Section 2.14 concludes. For the benefit of the readers already familiar with factor regression, we remark that our key methodological contributions are in Sections 2.7 - 2.12.

2.2 Bayesian factor analysis

Factor analysis (FA) models describe the observations $\mathsf{x}_i = (x_{i1}, x_{i2}, \ldots, x_{ip})^\top \in \mathbb{R}^p$, for $i = 1, \ldots, n$ individuals as a regression over latent variables $\mathsf{z}_i \in \mathbb{R}^q$, called latent coordinates or factors. Those latent coordinates tend to have a low-dimension $q \ll p$, making them easier to interpret and visualise. Let X be the $n \times p$ matrix with the i^{th} row equal to $\mathsf{x}_i^\top$ and Z the $n \times q$ matrix of latent coordinates, containing $\mathsf{z}_i^\top$ in the i^{th} row. More formally, the factor analysis model is defined as

$$\mathsf{x}_i = M\mathsf{z}_i + \mathsf{e}_i, \tag{2.1}$$

where $M \in \mathbb{R}^{p \times q}$ is the matrix of factor loadings and $\mathsf{e}_i \in \mathbb{R}^p$ is the error, distributed as $\mathsf{e}_i \sim N(0, \mathcal{T}^{-1})$ independently across $i = 1, \ldots, n$, with $\mathcal{T}^{-1}$ a diagonal matrix. Factors are assumed to be standard normal, $\mathsf{z}_i \sim N(0, \mathbf{I})$, independent across $i = 1, \ldots, n$ and also independent of e_i.

Without lost of generality, we have assumed that observations x_i have been mean centred, through this thesis. The non-centred model

$$\mathsf{x}_i = \mu + M\mathsf{z}_i + \mathsf{e}_i \tag{2.2}$$

with $\mu \in \mathbb{R}^p$, can be seen as the centred model $\widetilde{\mathsf{x}}_i = M\mathsf{z}_i + \mathsf{e}_i$ with $\widetilde{\mathsf{x}} = \mathsf{x}_i - \mu$ and estimating the mean with the sample mean $\widehat{\mu} = \bar{X} = \frac{1}{n}\sum_i^n \mathsf{x}_i$.

Alternatively, Equation (2.1) can be given in matrix notation as

$$X = ZM^\top + E, \tag{2.3}$$

with E $n \times p$ matrix of errors, containing $\mathsf{e}_i^\top$ in the i^{th} row.

Integrating out the factors, the implied marginal density of x_i is $f(\mathsf{x}_i \mid M, \mathcal{T}) = N(0, MM^\top + \mathcal{T}^{-1})$. Then, the covariance structure $\text{Cov}[\mathsf{x}_i \mid M, \mathcal{T}] = MM^\top + \mathcal{T}^{-1}$ can

be decomposed with at most $pq + p$ parameters instead of $p + p(p-1)/2 = p(p+1)/2$.

The factor model is non-identifiable up to orthogonal transformations, of the form $M^{*\top} = A^\top M^\top$ and $Z^* = ZA$, where A is any orthogonal $q \times q$ matrix. That is, the factor model in (2.3) can equivalently be rewritten as

$$X = Z^* M^{*\top} + E. \tag{2.4}$$

Clearly, both factor models generate the same covariance structure

$$\text{Cov}[\mathsf{x}_i \mid M, \mathcal{T}, A] = MM^\top + \mathcal{T}^{-1} = MAA^\top M^\top + \mathcal{T}^{-1} = M^* M^{*\top} + \mathcal{T}^{-1} \tag{2.5}$$

To obtain unique point estimates of M and Z, several strategies have been developed. One option is restricting the parameter space. Seber (1984) constrained M such that $M^\top AA^\top M$ is diagonal. Lopes and West (2004) restricted M to be lower-triangular with a strictly positive diagonal, $m_{jj} > 0$, and assumed M to be full-rank. More recently, Frühwirth-Schnatter and Lopes (2018) suggested a factor reordering via a Generalized Lower Triangular loading matrix. However, under this approach the interpretation of M depends on the arbitrary ordering of the columns in X, and it gives special roles to the first factors. Another option is to encourage sparsity in M, e.g. the classical varimax solution (Kaiser, 1958) maximises the variance in the squared rotated loadings. A more modern strategy is to favour sparse solutions containing exact zero loadings, e.g. Ročková and George (2017) proposed an EM algorithm that seeks rotations based on a so-called Parameter Expansion that aims to avoid local suboptimal regions. We adopt a similar strategy where sparse solutions are preferred by the introduced non-local penalties. This prior formulation will be discussed in Section 2.4.

2.2.1 Inference methods

Several parameter estimation methods have been developed by others to infer the loadings M and precision $\mathcal{T}$. We outline the principal component solution of factor analysis and the Maximum likelihood methods; in subsequent sections we discuss Bayesian solutions

Principal component solution of factor analysis

At the core of this method is the fact that, given the number of latent factors q, the sample covariance matrix $S = \frac{1}{n} X^\top X$ can be approximated by

$$\frac{1}{n} X^\top X \approx \widehat{M}\widehat{M}^\top + \widehat{\mathcal{T}}^{-1}. \tag{2.6}$$

Thus, the estimation of M and $\mathcal{T}$ is given by the following two-step approach:

Step 1: Consider the eigendecomposition of $\frac{1}{n}X^\top X$ where $l_1 \leq l_2 \leq \cdots \leq l_q$ are the eigenvalues and $u_1, \ldots, u_q$ are the eigen vectors. Then set

$$\widehat{M} = [\sqrt{l_1}u_1, \ldots \sqrt{l_q}u_q]. \tag{2.7}$$

Step 2: Typically one takes the precision as

$$\widehat{\tau}_{jj}^{-1} = \max\{0, S_{jj} - \widehat{m}_j^\top \widehat{m}_j\}. \tag{2.8}$$

where $S = \frac{1}{n}X^\top X$ and S_{jj} its (j, j) element.

Maximum likelihood method

Recall that $\mathrm{x}_i \mid M, \mathcal{T} \sim N(0, MM^\top + \mathcal{T}^{-1})$. We then aim to maximise the log-likelihood

$$\log \mathrm{p}(X \mid M, \mathcal{T}) \propto -\frac{1}{2}\sum_{i=1}^{n}\left[\mathrm{x}_i^\top (MM^\top + \mathcal{T}^{-1})^{-1}\mathrm{x}_i\right] - \frac{n}{2}\log|MM^\top + \mathcal{T}^{-1}|. \tag{2.9}$$

The estimators of $\widehat{M}$ and $\widehat{\mathcal{T}}^{-1}$ do not have a closed form, hence a numerical optimisation scheme – often Expectation-Maximisation (EM, Dempster et al. (1977)) – is normally used to obtain them. When incorporating priors to the FA model and performing posterior inference, one might also use EM algorithms, which give posterior modes. In addition to EM, estimation can also be carried out via MCMC algorithms (Lopes and West, 2004) to obtain the full posterior, or approximated via variational inference (Ghahramani and Beal, 2000). In this thesis we will provide deterministic optimisations to maximise the log-posterior via an Expectation-Maximisation algorithm that proved to be computationally efficient.

2.2.2 Prior formulation

To complete the Bayesian model specification we set some priors. As a first step we consider an improper flat prior on the loadings

$$\mathrm{p}(M) \propto 1. \tag{2.10}$$

An obvious limitation of (2.10) is that it does not induce any shrinkage or sparsity, we defer such extensions to Section 2.4.

Prior specification is then completed with a prior on diagonal elements of $\mathcal{T}$. We

assume independent gamma priors

$$\tau_j \sim \text{Gamma}(\eta/2, \eta\xi/2) \tag{2.11}$$

$j = 1, \ldots, p$. By default we set $\eta = \xi = 1$, this choice of hyper-parameters lead to relatively diffuse but proper priors. Figure 4 provides a Directed acyclic graph (DAG) for our model.

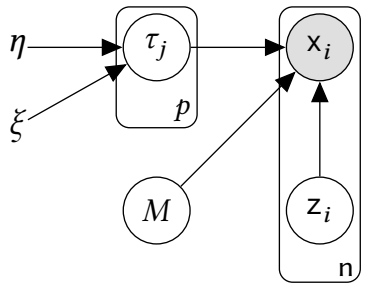


Figure 4. DAG for Bayesian factor analysis for Flat loadings matrix.

2.2.3 EM algorithm for factor analysis model under uniform p(M)

The Expectation-Maximisation (EM) algorithm aims to maximise the log-posterior $\log \text{p}(M, \mathcal{T} \mid X)$ by working with the complete-data log-posterior $\log \text{p}(M, \mathcal{T} \mid X, Z)$. This algorithm has two steps: the E-step calculates the expected log-likelihood w.r.t $\text{p}(Z \mid \widehat{M}, \widehat{\mathcal{T}}, X)$, where $\widehat{M}$ and $\widehat{\mathcal{T}}$ are the current values of M and $\mathcal{T}$. The M-step maximises $\mathbb{E}[\log \text{p}(M, \mathcal{T} \mid X, Z)]$ w.r.t. M and $\mathcal{T}$ giving a new update for them.

To ease the notation, let $\widehat{\Delta} = (\widehat{M}, \widehat{\mathcal{T}})$ be the current values of M and $\mathcal{T}$

Algorithm 1: EM algorithm for factor analysis model with uniform p(M)

initialise $\widehat{M} = M^{(0)}, \widehat{\mathcal{T}} = \mathcal{T}^{(0)}$
while $\varepsilon > \varepsilon^*, \varepsilon_M > \varepsilon_M^*$ *and* $t < T$ **do**
  E-step:
    Latent factors: $\mathbb{E}[\mathsf{z}_i | \widehat{\Delta}, X] = (\mathbf{I}_q + \widehat{M}^\top \widehat{\mathcal{T}} \widehat{M})^{-1} \widehat{M}^\top \widehat{\mathcal{T}} \mathsf{x}_i$
  M-step:
    Loadings: $\widehat{M} = \left[\sum_{i=1}^n \mathsf{x}_i \mathbb{E}[\mathsf{z}_i^\top \mid \widehat{\Delta}, X]\right] \left[\sum_{i=1}^n \mathbb{E}[\mathsf{z}_i \mathsf{z}_i^\top \mid \widehat{\Delta}, X]\right]^{-1}$
    Variances: $\widehat{\mathcal{T}}^{-1} = \frac{1}{n+\eta-2} \text{diag}\left\{\sum_{i=1}^n \left(\mathsf{x}_i \mathsf{x}_i^\top - \mathsf{x}_i \mathbb{E}[\mathsf{z}_i \mid \widehat{\Delta}, X]^\top \widehat{M}^\top\right) + \eta\xi \mathbf{I}_p\right\}$
  set $\Delta^{(t+1)} = \widehat{\Delta}$ and $M^{(t+1)} = \widehat{M}$
  compute $\varepsilon = Q(\Delta^{t+1}) - Q(\Delta^t)$, $\varepsilon_M = \max_{j,k} ||m_{jk}^{(t+1)} - m_{jk}^{(t)}||$ and $t = t + 1$
end

The E-step takes the expectation

$$
\begin{aligned}
Q(\Delta) =& \mathbb{E}_{Z|\widehat{\Delta},X}[\log \mathrm{p}(M, \mathcal{T} \mid X, Z)] \propto \mathbb{E}_{Z|\widehat{\Delta},X}\left[\log \mathrm{p}(X, Z \mid M, \mathcal{T}) + \log \mathrm{p}(M, \mathcal{T})\right] \\
&\propto -\frac{1}{2}\sum_{i=1}^{n} \mathbb{E}_{Z|\widehat{\Delta},X}\left[(\mathsf{x}_i - M\mathsf{z}_i)^\top \mathcal{T}(\mathsf{x}_i - M\mathsf{z}_i)\right] + \frac{n}{2}\log|\mathcal{T}| \\
&\quad + \sum_{j=1}^{p}\left[\frac{\eta - 2}{2}\log(\tau_j) - \frac{\eta\xi}{2}\mathcal{T}_j\right] \\
&= -\frac{1}{2}\sum_{i=1}^{n} \mathbb{E}_{Z|\widehat{\Delta},X}\left[\mathsf{x}_i^\top \mathcal{T}\mathsf{x}_i - 2\mathsf{x}_i\mathcal{T}M\mathsf{z}_i + \mathsf{z}_i^\top M^\top \mathcal{T}M\mathsf{z}_i\right] \\
&\quad + \frac{n + \eta - 2}{2} log|\mathcal{T}| - \frac{\eta\xi}{2}\operatorname{tr}(\mathcal{T}) \\
&= -\frac{1}{2}\sum_{i=1}^{n}\left[\mathsf{x}_i^\top \mathcal{T}\mathsf{x}_i - 2\mathsf{x}_i\mathcal{T}M\mathbb{E}[\mathsf{z}_i \mid \widehat{\Delta}, X] + \operatorname{tr}(M^\top \mathcal{T}M\mathbb{E}[\mathsf{z}_i\mathsf{z}_i^\top \mid \widehat{\Delta}, X])\right] \\
&\quad + \frac{n + \eta - 2}{2} log|\mathcal{T}| - \frac{\eta\xi}{2}\operatorname{tr}(\mathcal{T}),
\end{aligned}
\tag{2.12}
$$

where $\operatorname{tr}(A)$ is the trace of matrix A.

Note that Expression (2.12) only depends on Z through the conditional posterior mean

$$
\mathbb{E}[\mathsf{z}_i \mid \widehat{\Delta}, X] = (\mathbf{I}_q + \widehat{M}^\top \widehat{\mathcal{T}}\widehat{M})^{-1}\widehat{M}^\top \widehat{\mathcal{T}}\mathsf{x}_i \tag{2.13}
$$

and the conditional second moments

$$
\mathbb{E}[\mathsf{z}_i\mathsf{z}_i^\top \mid \widehat{\Delta}, X] = (\mathbf{I}_q + \widehat{M}^\top \widehat{\mathcal{T}}\widehat{M})^{-1} + \mathbb{E}[\mathsf{z}_i \mid \widehat{\Delta}, X]\mathbb{E}[\mathsf{z}_i \mid \widehat{\Delta}, X]^\top. \tag{2.14}
$$

The M-step consists in maximising Equation (2.12) with respect to Δ. To this end, we set its partial derivatives to 0, as shown below.

$$
\frac{\partial Q}{\partial M} = -\frac{1}{2}\sum_{i=1}^{n}\left[-2\widehat{\mathcal{T}}\mathsf{x}_i\mathbb{E}[\mathsf{z}_i^\top \mid \widehat{\Delta}, X] + 2\widehat{\mathcal{T}}\widehat{M}\mathbb{E}[\mathsf{z}_i\mathsf{z}_i^\top \mid \widehat{\Delta}, X]\right] = 0 \tag{2.15}
$$

The maximum of the loadings M can be found solving (2.15) as:

$$
\widehat{M} = \left[\sum_{i=1}^{n}\mathsf{x}_i\mathbb{E}[\mathsf{z}_i^\top \mid \widehat{\Delta}, X]\right]\left[\sum_{i=1}^{n}\mathbb{E}[\mathsf{z}_i\mathsf{z}_i^\top \mid \widehat{\Delta}, X]\right]^{-1} \tag{2.16}
$$

Analogously, maximisation of $\mathcal{T}$ is obtained by taking the derivative

$$\begin{aligned}\frac{\partial Q}{\partial \mathcal{T}} = &-\frac{1}{2}\sum_{i=1}^{n}\left[\mathsf{x}_i\mathsf{x}_i^\top - 2\mathsf{x}_i\mathbb{E}[\mathsf{z}_i \mid \widehat{\Delta}, X]^\top\widehat{M}^\top + \widehat{M}\mathbb{E}[\mathsf{z}_i\mathsf{z}_i^\top \mid \widehat{\Delta}, X]\widehat{M}^\top\right] \\ &+ \frac{n+\eta-2}{2}\widehat{\mathcal{T}}^{-1} - \frac{\eta\xi}{2}\mathbf{I}_p = 0.\end{aligned} \tag{2.17}$$

Substituting Equation (2.15) and using the diagonal constraint we obtain:

$$\widehat{\mathcal{T}}^{-1} = \frac{1}{n+\eta-2}\operatorname{diag}\left\{\sum_{i=1}^{n}\left(\mathsf{x}_i\mathsf{x}_i^\top - \mathsf{x}_i\mathbb{E}[\mathsf{z}_i \mid \widehat{\Delta}, X]^\top\widehat{M}^\top\right) + \eta\xi\mathbf{I}_p\right\} \tag{2.18}$$

Algorithm 1 summarises the EM algorithm. The stopping criterion is reaching a tolerance ε^* in the log-posterior change, a maximum number of iterations T or a change ε_M^* on the loadings. By default we set $\varepsilon^* = 0.001$, $T = 100$ and $\varepsilon_M^* = 0.05$. The EM algorithm increases the complete-data log-posterior at each iteration, however it does not guarantee convergence to a global maximum. Thus, initial values are crucial in order to obtain a good performance. Parameter initialisation is discussed in Section 2.11.

2.3 Latent factor cardinality q

The choice of the cardinality of the latent factors q is a crucial aspect in FA. Until now q was assumed known. In practice, there are several strategies to infer q. One option is to treat the problem as a model selection, choosing q with the smallest Akaike information criterion (AIC) or Bayesian information criterion (BIC). Another option is to consider a single model with large q and set penalties or priors that induce sparse solutions, where only some small proportion of the loadings are non-zero, easing the interpretation of the model. Some recent strategies include a LASSO-based method (Witten et al., 2009), horseshoe priors (Carvalho et al., 2009), an Indian buffet process (Knowles and Ghahramani, 2011), and an infinite factor model (Dunson and Bhattacharya, 2011) among others. In this thesis we focus on continuous mixture penalties that build on the approach by Ročková and George (2014, 2017).

2.4 Spike-and-Slab prior

A traditional Bayesian approach to variable selection is the spike-and-slab prior, a two-component mixture prior (Mitchell and Beauchamp, 1988; George and McCulloch, 1993). This prior aims to discriminate those loadings that warrant inclusion, modelled by the

slab component, from those that should be excluded, modelled by the spike component.

Specifically, as mention in Equation (1.15), a spike-and-slab prior density for the loadings M has the form

$$\mathrm{p}(M \mid \gamma, \lambda_0, \lambda_1) = \prod_{j=1}^{p} \prod_{k=1}^{q} (1 - \gamma_{jk}) \mathrm{p}(m_{jk} \mid \lambda_0, \gamma_{jk} = 0) + \gamma_{jk} \mathrm{p}(m_{jk} \mid \lambda_1, \gamma_{jk} = 1), \quad (2.19)$$

where $\mathrm{p}(m_{jk} \mid \lambda_0, \gamma_{jk} = 0)$ is a continuous density, λ_0 is a given dispersion parameter of the spike component and $\lambda_1 > \lambda_0$ is that of the slab component. The indicators $\gamma_{jk} \in \{0, 1\}$ signal which m_{jk} were generated by each component, and serve as a proxy for which loadings are significantly non-zero.

Through this thesis, we consider a hierarchical prior over the latent indicator

$$\gamma = \{\gamma_{jk}, j = 1, \ldots, p, k = 1, \ldots, q\}$$

as follows,

$$\begin{aligned} \gamma_{jk} \mid \zeta_k &\sim \text{Bernoulli}(\zeta_k), \\ \zeta_k \mid a_\zeta, b_\zeta &\sim \text{Beta}\left(\frac{a_\zeta}{k}, b_\zeta\right), \end{aligned} \quad (2.20)$$

with independence across (j, k) where $a_\zeta > 0$ and $b_\zeta > 0$ are given prior parameters.

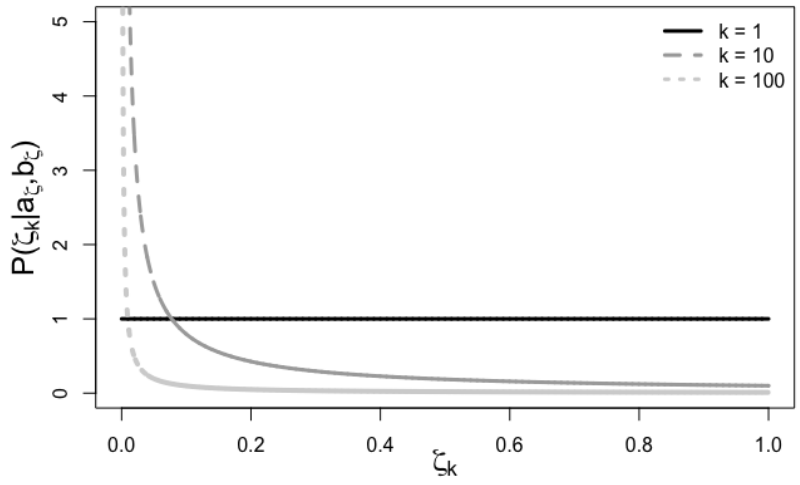


Figure 5. Comparison of Beta $\left(\frac{1}{k}, 1\right)$ at different values of k.

By default we set $a_\zeta = b_\zeta = 1$, which leads to a uniform prior for the first factor ($k = 1$), $\zeta_k \mid a_\zeta, b_\zeta \sim \mathrm{U}(0, 1)$. Furthermore, note that $\frac{a_\zeta}{k}$ encourages increasingly sparse solutions in subsequent factors. That is, related to our earlier discussion of non-identifiability (Section 2.2), we encourage loadings where the first factors have larger importance, leading to solutions that are sparse both in the rank of M and its non-zero entries. Figure 6 presents a DAG of the spike-and-slab prior

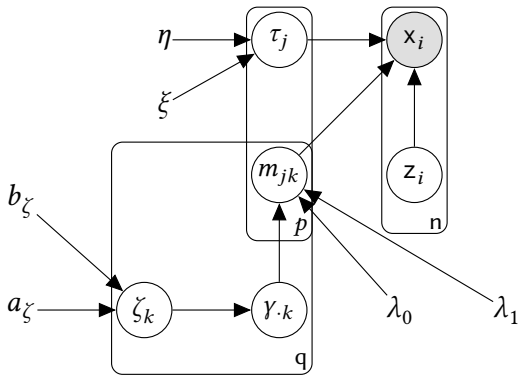


Figure 6. Directed acyclic graph (DAG) for Bayesian factor analysis with batch effect correction for Spike-and-slab prior on the loadings matrix.

2.5 Normal-spike-and-slab

2.5.1 Formulation

We first describe the Normal-spike-and-slab prior by George and McCulloch (1993) were the spike is a Normal density with a small variance λ_0 and the slab a Normal distribution with large variance λ_1. The Normal-spike-and-slab is

$$\mathrm{p}(m_{jk} \mid \gamma_{jk} = l, \lambda_l) = N(m_{jk}; 0, \lambda_l), \tag{2.21}$$

with $l = \{0, 1\}$. The continuity of the spike distribution gives closed form expressions for the EM algorithm, making it computationally appealing. We refer to (2.21) as Normal-SS.

2.5.2 EM algorithm for Normal-SS

Akin to the Flat prior, we outline an EM algorithm to infer $\Delta = (M, \mathcal{T}, \zeta)$. Algorithm 2 summarises this maximisation.

The E-step We first take the expectation of the complete-data log-posterior with re-

Algorithm 2: EM algorithm for factor analysis model with spike-and-slab p(M)

initialise $\widehat{M} = M^{(0)}$, $\widehat{\mathcal{T}} = \mathcal{T}^{(0)}$, $\widehat{\zeta} = \zeta^{(0)}$
while $\varepsilon > \varepsilon^*$, $\varepsilon_M > \varepsilon_M^*$ *and* $t < T$ **do**
E-step:
Latent factors: $\mathbb{E}[\mathsf{z}_i|\widehat{\Delta}, X] = (\mathbf{I}_q + \widehat{M}^\top\widehat{\mathcal{T}}\widehat{M})^{-1}\widehat{M}^\top\widehat{\mathcal{T}}\mathsf{x}_i$
Latent indicators$^+$: $\mathbb{E}[\gamma_{jk} \mid \widehat{\Delta}] = \widehat{p}_{jk}$
M-step:
Loadings$^+$: $\widehat{m}_{jk} = \arg\max_{m_{jk}} Q_1(\widehat{\Delta})$
Precision: $\widehat{\mathcal{T}}^{-1} = \frac{1}{n+\eta-2}\operatorname{diag}\left\{\sum_i \left(\mathsf{x}_i\mathsf{x}_i^\top - 2\mathsf{x}_i\mathbb{E}[\mathsf{z}_i \mid \widehat{\Delta}, X]^\top\widehat{M}^\top + \widehat{M}\mathbb{E}[\mathsf{z}_i\mathsf{z}_i^\top \mid \widehat{\Delta}, X]\widehat{M}^\top\right) + \eta\xi\mathbf{I}_p\right\}$
Weights: $\widehat{\zeta}_k = \frac{\sum_{j=1}^p \widehat{p}_{jk} + \frac{a_\zeta}{k} - 1}{\frac{a_\zeta}{k} + b_\zeta + p - 1}$
set $\Delta^{(t+1)} = \widehat{\Delta}$ and $M^{(t+1)} = \widehat{M}$
compute $\varepsilon = Q(\Delta^{t+1}) - Q(\Delta^t)$, $\varepsilon_M = \max||m_{jk}^{(t+1)} - m_{jk}^{(t)}||$ and $t = t + 1$
end
$^+$ see Sections 2.5.2, 2.6.2, 2.8.2 and 2.9.2 for details.

spect to the latent variables and conditioning upon the current $\widehat{\Delta}$:

$$Q(\Delta) \propto \mathbb{E}_{z,\gamma|\widehat{\Delta},X}\left[\log \mathrm{p}(X, Z, \gamma \mid M, \mathcal{T}, \zeta) + \log \mathrm{p}(M, \mathcal{T}, \zeta)\right] \tag{2.22}$$

Due to the conjugate Normal-SS hierarchical construction, Expression (2.22) can be split in order to simplify the EM algorithm as $Q(\Delta) = C + Q_1(M, \mathcal{T}) + Q_2(\zeta)$, where:

$$\begin{aligned} Q_1(M, \mathcal{T}) = &-\frac{1}{2}\sum_{i=1}^n \left[\mathsf{x}_i^\top\mathcal{T}\mathsf{x}_i - 2\mathsf{x}_i^\top\mathcal{T}M\mathbb{E}[\mathsf{z}_i \mid \widehat{\Delta}, X] + \operatorname{tr}\left(M^\top\mathcal{T}M\mathbb{E}[\mathsf{z}_i\mathsf{z}_i^\top \mid \widehat{\Delta}, X]\right)\right] \\ &+ \frac{n+\eta-2}{2}\log \mid \mathcal{T} \mid - \frac{\eta\xi}{2}\operatorname{tr}(\mathcal{T}) \\ &- \frac{1}{2}\sum_{j=1}^p\sum_{k=1}^q m_{jk}^2\mathbb{E}\left[\frac{1}{(1-\gamma_{jk})\lambda_0 + \gamma_{jk}\lambda_1} \mid \widehat{\Delta}\right], \end{aligned} \tag{2.23}$$

$$Q_2(\zeta) = \sum_{j=1}^p\sum_{k=1}^q \log\left(\frac{\zeta_k}{1-\zeta_k}\right)\widehat{p}_{jk} + \sum_{k=1}^q\left((\frac{a_\zeta}{k} - 1)\log(\zeta_k) + (p + b_\zeta - 1)\log(1-\zeta_k)\right), \tag{2.24}$$

where $\widehat{p}_{jk} = \mathbb{E}[\gamma_{jk} \mid \widehat{\Delta}] = \mathrm{p}(\gamma_{jk} = 1 \mid \widehat{\Delta})$.

Expression (2.23) resembles the one for the flat prior in Section 2.2.3, plus an extra conditional expectation:

$$\mathbb{E}\left[\frac{1}{(1-\gamma_{jk})\lambda_0 + \gamma_{jk}\lambda_1} \mid \widehat{\Delta}\right] = \frac{1-\widehat{p}_{jk}}{\lambda_0} + \frac{\widehat{p}_{jk}}{\lambda_1},$$

with $\widehat{p}_{jk} = \mathrm{p}(\gamma_{jk} = 1 \mid \widehat{\Delta})$ given by

$$\begin{aligned}\widehat{p}_{jk} =& \frac{\mathrm{p}(\widehat{m}_{jk} \mid \gamma_{jk} = 1, \lambda_1)\mathrm{p}(\gamma_{jk} = 1)}{\mathrm{p}(\widehat{m}_{jk} \mid \gamma_{jk} = 0, \lambda_0)\mathrm{p}(\gamma_{jk} = 0) + \mathrm{p}(\widehat{m}_{jk} \mid \gamma_{jk} = 1, \lambda_1)\mathrm{p}(\gamma_{jk} = 1)} \\ =& \frac{1}{1 + \sqrt{\frac{\lambda_1}{\lambda_0}} \exp\left(-\frac{1}{2}\widehat{m}_{jk}^2 \left(\frac{1}{\lambda_0} - \frac{1}{\lambda_1}\right)\right) \frac{1-\mathbb{E}[\zeta_j]}{\mathbb{E}[\zeta_j]}}.\end{aligned} \tag{2.25}$$

Equation (2.25) is analogous to the EM posterior update for m_{jk} in a two-component Gaussian mixture (Ročková and George, 2014).

The first and second moments $\mathbb{E}[\mathrm{z}_i \mid \widehat{\Delta}, X]$ and $\mathbb{E}[\mathrm{z}_i\mathrm{z}_i^\top \mid \widehat{\Delta}, X]$ respectively are given in Equations (2.13) and (2.14) respectively.

The M-step. We proceed by optimising Q_1 and Q_2 independently, in 2 steps: a maximisation of Q_1 with respect to M and $\mathcal{T}$, followed by a maximisation of Q_2 with respect to ζ. Setting to 0 the partial derivatives with respect to M gives:

$$\frac{\partial Q}{\partial M} = -\frac{1}{2}\sum_{i=1}^{n}\left[-2\widehat{\mathcal{T}}\mathrm{x}_i\mathbb{E}[\mathrm{z}_i^\top \mid \widehat{\Delta}, X] + 2\widehat{\mathcal{T}}\widehat{M}\mathbb{E}[\mathrm{z}_i\mathrm{z}_i^\top \mid \widehat{\Delta}, X]\right] - \widehat{M} \circ \mathbb{E}[D_\gamma \mid \widehat{\Delta}] = 0, \tag{2.26}$$

with $D_\gamma \in \mathbb{R}^{p\times q}$, $d_{jk} = \frac{1}{(1-\gamma_{jk})\lambda_0 + \gamma_{jk}\lambda_1}$ and $A \circ B$ being the Hadamard (element-wise) product of two matrices A and B. Taking the j^{th} row of matrix M and solving equation (2.26) we obtain:

$$\widehat{m}_j = \left[\sum_{i=1}^{n}\left(\widehat{\tau}_j x_{ij}\mathbb{E}[\mathrm{z}_i^\top \mid \widehat{\Delta}, X]\right)\right]\left[\mathrm{diag}\{\mathbb{E}[d_{j1} \mid \widehat{\Delta}], \ldots, \mathbb{E}[d_{jq} \mid \widehat{\Delta}]\} + \sum_{i=1}^{n}\left(\widehat{\tau}_j\mathbb{E}[\mathrm{z}_i\mathrm{z}_i^\top \mid \widehat{\Delta}, X]\right)\right]^{-1}, \tag{2.27}$$

for $j = 1, \ldots, p$.

The partial derivative w.r.t. $\mathcal{T}$ is the same as Equation (2.17). However the new update for the loadings $\widehat{M}$ leads to a different solution, namely

$$\widehat{\mathcal{T}}^{-1} = \frac{1}{n+\eta-2}\mathrm{diag}\left\{\sum_{i=1}^{n}\left(\mathrm{x}_i\mathrm{x}_i^\top - 2\mathrm{x}_i\mathbb{E}[\mathrm{z}_i \mid \widehat{\Delta}, X]^\top\widehat{M}^\top + \widehat{M}\mathbb{E}[\mathrm{z}_i\mathrm{z}_i^\top \mid \widehat{\Delta}, X]\widehat{M}^\top\right) + \eta\xi\mathbf{I}_p\right\}. \tag{2.28}$$

Finally,

$$\frac{\partial Q_2}{\partial \zeta_k} = \frac{\sum_{j=1}^{p}\widehat{p}_{jk}}{\widehat{\zeta}_k - \widehat{\zeta}_k^2} + \frac{\frac{a_\zeta}{k} - 1}{\widehat{\zeta}_k} - \frac{p + b_\zeta - 1}{1 - \widehat{\zeta}_k} = 0. \tag{2.29}$$

Solving Equation (2.29):

$$\widehat{\zeta}_k = \frac{\sum_{j=1}^{p} \widehat{p}_{jk} + \frac{a_\zeta}{k} - 1}{\frac{a_\zeta}{k} + b_\zeta + p - 1}. \tag{2.30}$$

2.6 Laplace-spike-and-slab

2.6.1 Formulation

The Normal-SS prior does not give exact zeroes for the loadings $\widehat{M}$. We now formulate the Laplace-spike-and-slab, (Laplace-SS). Laplace-SS was introduced by Ročková and George (2018) and is a two-component mixture of double exponentials that shrinks small values of the loadings to exact zeroes. Its heavy tails and continuity make them appealing for FA, providing closed-form updates for the EM algorithm. Laplace-SS is of the form

$$\mathrm{p}(m_{jk} \mid \gamma_{jk}, \lambda_0, \lambda_1) = (1 - \gamma_{jk})\mathrm{Laplace}(m_{jk}; 0, \lambda_0) + \gamma_{jk}\mathrm{Laplace}(m_{jk}; 0, \lambda_1), \tag{2.31}$$

with a slab component with variance $2\lambda_0^2$, and a spike component with $2\lambda_1^2$, where

$$\mathrm{Laplace}(m_{jk}; 0, \lambda) = \frac{1}{2\lambda} \exp\left(\frac{- \mid m_{jk} \mid}{\lambda}\right).$$

2.6.2 EM algorithm for Laplace-SS

The E-step The expected complete-data log-posterior is the sum of two components. The first component is

$$\begin{aligned} Q_1(M, \mathcal{T}) = &- \frac{1}{2} \sum_{i=1}^{n} \left[\mathrm{x}_i^\top \mathcal{T} \mathrm{x}_i - 2\mathrm{x}_i^\top \mathcal{T} M \mathbb{E}[\mathrm{z}_i \mid \widehat{\Delta}, X] + \mathrm{tr}\left(M^\top \mathcal{T} M \mathbb{E}[\mathrm{z}_i \mathrm{z}_i^\top \mid \widehat{\Delta}, X] \right) \right] \\ &+ \frac{n + \eta - 2}{2} \log |\mathcal{T}| - \frac{\eta \xi}{2} \mathrm{tr}(\mathcal{T}) - \sum_{j=1}^{p} \sum_{k=1}^{q} |m_{jk}| \mathbb{E}\left[\frac{1 - \gamma_{jk}}{\lambda_0} + \frac{\gamma_{jk}}{\lambda_1} \mid \widehat{\Delta} \right]. \end{aligned} \tag{2.32}$$

and Q_2 has the same form as (2.24).

For (2.32) we set

$$\mathbb{E}\left[\frac{1 - \gamma_{jk}}{\lambda_0} + \frac{\gamma_{jk}}{\lambda_1} \mid \widehat{\Delta} \right] = \frac{1 - \widehat{p}_{jk}}{\lambda_0} + \frac{\widehat{p}_{jk}}{\lambda_1},$$

with $\mathbb{E}[\mathrm{z}_i \mid \widehat{\Delta}, X]$ and $\mathbb{E}[\mathrm{z}_i \mathrm{z}_i^\top \mid \widehat{\Delta}, X]$ as in (2.13) and (2.14).

The conditional expectation for Q_2, $\mathbb{E}[\gamma_{jk} \mid \widehat{\Delta}] = \mathrm{p}[\gamma_{jk} = 1 \mid m_{jk}] = \widehat{p}_{jk}$ is

$$\begin{aligned}\widehat{p}_{jk} =& \frac{\mathrm{p}(\widehat{m}_{jk} \mid \gamma_{jk} = 1, \lambda_1)\mathrm{p}(\gamma_{jk} = 1)}{\mathrm{p}(\widehat{m}_{jk} \mid \gamma_{jk} = 0, \lambda_0)\mathrm{p}(\gamma_{jk} = 0) + \mathrm{p}(\widehat{m}_{jk} \mid \gamma_{jk} = 1, \lambda_1)\mathrm{p}(\gamma_{jk} = 1)} \\ =& \frac{1}{1 + \frac{\lambda_1}{\lambda_0} \exp\left(-|\widehat{m}_{jk}| \left(\frac{1}{\lambda_0} - \frac{1}{\lambda_1}\right)\right) \frac{1-\mathbb{E}[\zeta_j]}{\mathbb{E}[\zeta_j]}}\end{aligned} \tag{2.33}$$

The M-step update for M is obtained by setting to 0 the partial derivatives which are defined for $m_{jk} \neq 0$, and considering separately the non-differentiability points $m_{jk} = 0$. For $m_{jk} \neq 0$ we have

$$\frac{\partial Q}{\partial M} = -\frac{1}{2} \sum_{i=1}^{n} \left[-2\mathcal{T}\mathsf{x}_i \mathbb{E}[\mathsf{z}_i^\top \mid \widehat{\Delta}, X] + 2\mathcal{T}M\mathbb{E}[\mathsf{z}_i \mathsf{z}_i^\top \mid \widehat{\Delta}, X]\right] - D^{\gamma, M} = 0, \tag{2.34}$$

with $D^{\gamma,M} \in \mathbb{R}^{p \times q}$ with element (j,k) being: $d_{jk}^{\gamma,M} = \mathrm{sign}(m_{jk})\mathbb{E}[d_{jk} \mid \widehat{\Delta}]$. To maximise (2.32), we consider a coordinate descent algorithm (CDA) that leads to closed-form updates. Viewing (2.34) with respect to m_{jk}, when $m_{jk} \neq 0$ we obtain:

$$\begin{aligned}\frac{\partial Q_1}{\partial m_{jk}} =& -\left(\sum_{i=1}^{n} \widehat{\tau}_{jj}\mathbb{E}[z_{ik} z_{ik}^\top \mid \widehat{\Delta}, X]\right) \widehat{m}_{jk} + \left(\sum_{i=1}^{n} \left[\widehat{\tau}_{jj} x_{ij} \mathbb{E}[z_{ik} \mid \widehat{\Delta}, X]\right.\right. \\ & \left.\left. - \sum_{r \neq k}^{q} \widehat{m}_{jr}\widehat{\tau}_{jj}\mathbb{E}[z_{ir} z_{ik}^\top \mid \widehat{\Delta}, X]\right] - \mathrm{sign}(\widehat{m}_{jk}) \left[\frac{1 - \widehat{p}_{jk}}{\lambda_0} + \frac{\widehat{p}_{jk}}{\lambda_1}\right]\right) \\ =& a\widehat{m}_{jk} + b + c \cdot \mathrm{sign}(\widehat{m}_{jk}) = 0\end{aligned} \tag{2.35}$$

for $j = 1, \ldots, p$, where (a, b, c) do not depende on m_{jk} and are given in the first line of Equation (2.35). The solution to (2.35) is then compared to m_{jk}. Lemma 2.1 summarises the global maximum for the loadings.

Lemma 2.1. *Let $f(m_{jk}) = \frac{a}{2}m_{jk}^2 + bm_{jk} + c|m_{jk}|$, where $a < 0$ and $c < 0$. Define $m_{jk}^+ = \frac{-(b+c)}{a}$ and $m_{jk}^- = \frac{-(b-c)}{a}$.*

If $b > -c$, then $m_{jk}^+ = \arg\max_{m_{jk}} f(m_{jk})$. If $b < c$, then $m_{jk}^- = \arg\max_{m_{jk}} f(m_{jk})$. If $c \leq b \leq -c$, then $0 = \arg\max_{m_{jk}} f(m_{jk})$.

Proof. Our purpose is to find the maximum of $f(m_{jk}) = \frac{a}{2}m_{jk}^2 + bm_{jk} + c|m_{jk}|$, where $a < 0$, and $c < 0$. Setting $\frac{\partial Q_1}{\partial m_{jk}} = 0$, we obtain

$$\frac{\partial Q_1}{\partial m_{jk}} = am_{jk} + b + c \cdot \mathrm{sign}(m_{jk}) = 0.$$

- For $m_{jk} > 0$, we look for the solutions of $am_{jk} + b + c = 0$. Note $a < 0$ and $c < 0$.

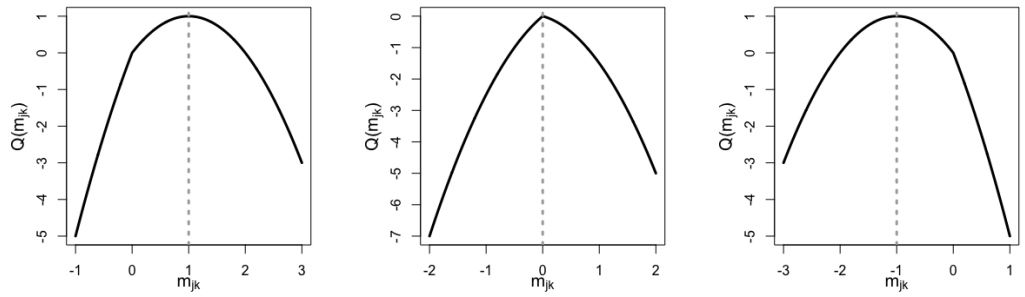
Thus

$$arg\max_{m_{jk}\geq 0} f(m_{jk}) = \begin{cases} m_{jk}^{+} := \frac{-(b+c)}{a} & b > -c \\ 0 & \text{otherwise} \end{cases}$$

- For $m_{jk} < 0$, we look for the solutions of $am_{jk} + b - c = 0$. Thus

$$arg\max_{m_{jk}\leq 0} f(m_{jk}) = \begin{cases} m_{jk}^{-} := \frac{-(b-c)}{a} & b < c \\ 0 & \text{otherwise} \end{cases} \quad \square$$

Figure 7 presents a visual representation of Q_1 in function of m_{jk} for different values of b. The updates for $\widehat{\mathcal{T}}$ and $\widehat{\zeta}_k$ are the ones given in Equations (2.28) and (2.30) respectively.



(A) $b > -c$ (B) $c < b < -c$ (C) $b < c$

Figure 7. Maximisation of m_{jk} for Laplace-SS.

A potential concern with Normal-SS and Laplace-SS is that the slab density assigns non-negligible probability to regions of the parameter space that are also consistent with the spike, namely when m_{jk} lies close to zero. We will address this via non-local priors and show that these, by enforcing separation between the two components, help increase sensitivity.

2.7 Non-local priors

Non-local priors (NLPs) are a family of distributions that, conditional on the alternative hypothesis, assign vanishing prior density to a neighbourhood of the null hypothesis (Johnson and Rossell, 2010). Definition 2.2 is an adaptation of the definition in Johnson and Rossell (2010) to (2.19).

Definition 2.2. An absolutely continuous measure with density $p(m_{jk}|\gamma_{jk} = 1)$ is a non-local prior if $\lim_{m_{jk}\to 0} p(m_{jk}|\gamma_{jk} = 1) = 0$.

We call any prior not satisfying Definition 2.2 a local prior. Non-local priors possess appealing properties for Bayesian model selection. They discard spurious parameters faster as the sample size n grows, but preserve exponential rates to detect important coefficients (Johnson and Rossell, 2010) and can lead to improved parameter estimation shrinkage (Rossell and Telesca, 2017). To illustrate the motivation for NLPs in our setting consider Figure 8. Normal-SS assigns positive probability to $m_{jk} = 0$. Correspondingly, the conditional inclusion probability $p(\gamma_{jk} = 1 \mid m_{jk})$ remains non-negligible, even when $m_{jk} = 0$ (lower left panel).

2.8 Normal-spike-and-MOM-slab

2.8.1 Formulation

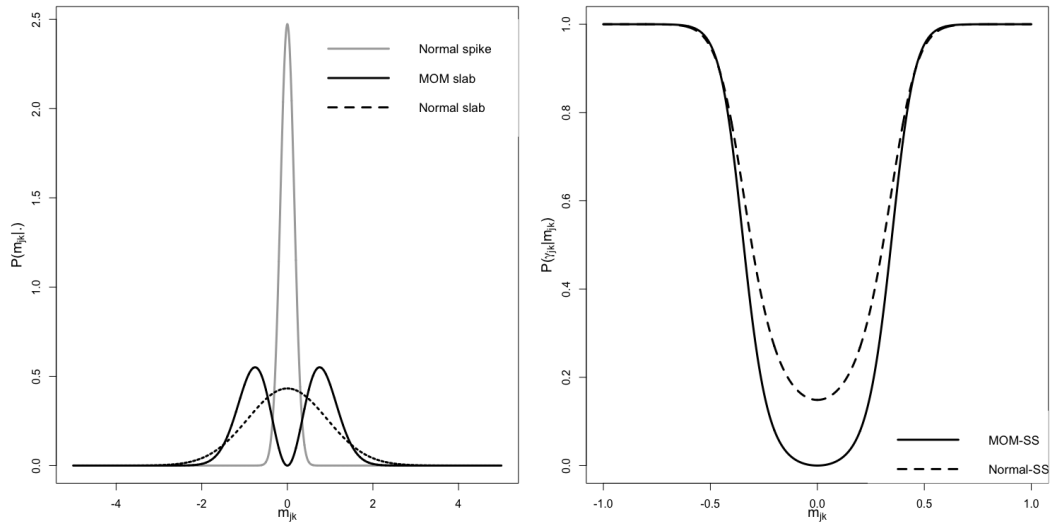


Figure 8. Prior comparison (left panel) for m_{jk} under Normal-SS and MOM-SS and its inclusion probabilities $p(\gamma_{jk} \mid m_{jk})$ (right panels). Scales (λ_0, λ_1) are set to the defaults from Section 2.10.

As an alternative to the Normal-SS, we consider a product moment (pMOM) prior (Johnson and Rossell, 2012).

$$\begin{aligned} p(m_{jk} \mid \gamma_{jk} = 0, \widetilde{\lambda}_0) &= N(m_{jk}; 0, \widetilde{\lambda}_0), \\ p(m_{jk} \mid \gamma_{jk} = 1, \widetilde{\lambda}_1) &= \frac{m_{jk}^2}{\widetilde{\lambda}_1} N(m_{jk}; 0, \widetilde{\lambda}_1). \end{aligned} \tag{2.36}$$

We refer to (2.36) as MOM-SS. This prior assigns zero density to $m_{jk} = 0$ given $\gamma_{jk} = 1$, which implies $\mathrm{p}(\gamma_{jk} = 1 \mid m_{jk} = 0) = 0$ (Figure 8). Prior elicitation for $\widetilde{\lambda}_0$ and $\widetilde{\lambda}_1$ is discussed in Section 2.10. From a computational point of view, the EM algorithm can accommodate this extension by using a trivial extra gradient evaluation at negligible additional cost relative to the Normal-SS. Parameter estimation and algebraic details are described in Section 2.8.2. The prior on the inclusion indicators is set as in Equation (2.20).

2.8.2 EM algorithm for MOM-SS

The E-step: Analogous to Local spike-and-slabs (L-SSs), we first take the expected complete-data log-posterior $Q(\Delta) = C + Q_1(\theta, M, \beta, \mathcal{T}_{\mathsf{b}_i}) + Q_2(\zeta)$. By construction Q_2 is of the same form than in Equation (2.24) and Q_1 is given by

$$\begin{aligned} Q_1(\theta, M, \beta, \mathcal{T}) = &- \frac{1}{2}\sum_{i=1}^{n}\left[\mathsf{x}_i^\top \mathcal{T}\mathsf{x}_i - 2\mathsf{x}_i^\top \mathcal{T} M \mathbb{E}[\mathsf{z}_i \mid \widehat{\Delta}, X] + \mathrm{tr}\left(M^\top \mathcal{T} M \mathbb{E}[\mathsf{z}_i\mathsf{z}_i^\top \mid \widehat{\Delta}, X]\right)\right] \\ &+ \frac{n+\eta-2}{2}\log \mid \mathcal{T} \mid - \frac{\eta\xi}{2}\mathrm{tr}(\mathcal{T}) \\ &- \frac{1}{2}\sum_{j=1}^{p}\sum_{k=1}^{q} m_{jk}^2 \mathbb{E}\left[d_{jk} \mid \widehat{\Delta}\right] + \sum_{j=1}^{p}\sum_{k=1}^{q}\mathbb{E}[\gamma_{jk} \mid \widehat{\Delta}]\log(m_{jk}^2). \end{aligned} \tag{2.37}$$

$\mathbb{E}[\mathsf{z}_i|\widehat{\Delta}, X]$ and $\mathbb{E}[\mathsf{z}_i\mathsf{z}_i^\top \mid \widehat{\Delta}, X]$ are the same as in Equations (2.13) and (2.14) respectively for the Flat priors. The new conditional expectation for the inclusion probability $\mathbb{E}[\gamma_{jk} \mid \widehat{\Delta}] = \widehat{p}_{jk}$ is

$$\widehat{p}_{jk} = \frac{1}{1 + \frac{\widetilde{\lambda}_1}{\widetilde{m}_{jk}^2}\sqrt{\frac{\widetilde{\lambda}_1}{\widetilde{\lambda}_0}}\exp\left(-\frac{1}{2}\widehat{m}_{jk}^2\left(\frac{1}{\widetilde{\lambda}_0} - \frac{1}{\widetilde{\lambda}_1}\right)\right)\frac{1-\mathbb{E}[\zeta_j]}{\mathbb{E}[\zeta_j]}} \tag{2.38}$$

and $\mathbb{E}[d_{jk} \mid \widehat{\Delta}] = \mathbb{E}\left[\frac{1}{(1-\gamma_{jk})\widetilde{\lambda}_0 + \gamma_{jk}\widetilde{\lambda}_1} \mid \widehat{\Delta}\right] = \frac{1-\widehat{p}_{jk}}{\widetilde{\lambda}_0} + \frac{\widehat{p}_{jk}}{\widetilde{\lambda}_1}$.

The M-step: We use a coordinate descent algorithm (CDA) that performs successive univariate optimisation on (2.37) with respect to each m_{jk}, in order to maximise the loadings. An advantage is that the updates have a closed-form that is computationally inexpensive. As a potential drawback it could require a larger number of iterations to converge relative to performing joint optimisation with respect to multiple elements in M. However, we have not found this to be a practical problem in our examples. See Sec-

tion 2.13. The partial derivative of (2.37) is:

$$\begin{aligned}\frac{\partial Q_1}{\partial M} = &\sum_{i=1}^{n}\left[\widehat{\mathcal{T}}\mathsf{x}_i\mathbb{E}[\mathsf{z}_i^\top \mid \widehat{\Delta}, X] - \widehat{\mathcal{T}}\widehat{M}\mathbb{E}[\mathsf{z}_i\mathsf{z}_i^\top \mid \widehat{\Delta}, X]\right] \\ &- \widehat{M}\circ\mathbb{E}[D_\gamma \mid \widehat{\Delta}] + 2\mathbb{E}[\gamma \mid \widehat{\Delta}]\circ\widehat{M}_{inv} = 0,\end{aligned} \tag{2.39}$$

with $D_\gamma \in \mathbb{R}^{p\times q}$, $d_{jk} = ((1-\gamma_{jk})\lambda_0 + \gamma_{jk}\lambda_1)^{-1}$, $\widehat{M}_{inv}$ a matrix with elements $1/\widehat{m}_{jk}$ and $A\circ B$ being the Hadamard (element-wise) product of two matrices A and B.

Viewing (2.39) with respect to m_{jk}:

$$\begin{aligned}\frac{\partial Q_1}{\partial m_{jk}} =& -\left(\mathbb{E}[d_{jk}] + \sum_{i=1}^{n}\widehat{\tau}_{jj}\mathbb{E}[z_{ik}z_{ik}^\top \mid \widehat{\Delta}, X]\right)\widehat{m}_{jk} + \left(\sum_{i=1}^{n}\left[\widehat{\tau}_{jj}x_{ij}\mathbb{E}[z_{ik} \mid \widehat{\Delta}, X]\right.\right. \\ &\left.\left. - \sum_{r\neq k}^{q}\widehat{m}_{jr}\widehat{\tau}_{jj}\mathbb{E}[z_{ir}z_{ik}^\top \mid \widehat{\Delta}, X]\right]\right) + \frac{2\mathbb{E}[\gamma_{jk}]}{\widehat{m}_{jk}} = 0 \\ \Leftrightarrow& -\left(\mathbb{E}[d_{jk}] + \sum_{i=1}^{n}\widehat{\tau}_{jj}\mathbb{E}[z_{ik}z_{ik}^\top \mid \widehat{\Delta}, X]\right)\widehat{m}_{jk}^2 + \left(\sum_{i=1}^{n}\left[\widehat{\tau}_{jj}x_{ij}\mathbb{E}[z_{ik} \mid \widehat{\Delta}, X]\right.\right. \\ &\left.\left. - \sum_{r\neq k}^{q}\widehat{m}_{jr}\widehat{\tau}_{jj}\mathbb{E}[z_{ir}z_{ik}^\top \mid \widehat{\Delta}, X]\right]\right)\widehat{m}_{jk} + 2\mathbb{E}[\gamma_{jk}] \\ =& a\widehat{m}_{jk}^2 + b\widehat{m}_{jk} + c = 0\end{aligned} \tag{2.40}$$

for $j = 1, \ldots, p$, where (a, b, c) do not depend on m_{jk} and are as given in (2.40). The global maximum of (2.40) is summarised in Lemma 2.3.

Lemma 2.3. *Let* $f(m_{jk}) = \frac{a}{2}m_{jk}^2 + bm_{jk} + \frac{c}{2}\log(m_{jk}^2)$, *where* $a < 0$ *and* $c > 0$. *Define* $\underline{m}_{jk} = \frac{-b-\sqrt{b^2-4ac}}{2a}$ *and* $\overline{m}_{jk} = \frac{-b+\sqrt{b^2-4ac}}{2a}$.
If $b > 0$, *then* $\underline{m}_{jk} = \arg\max_{m_{jk}} f(m_{jk})$. *If* $b < 0$, *then* $\overline{m}_{jk} = \arg\max_{m_{jk}} f(m_{jk})$. *If* $b = 0$, *then* $\overline{m}_{jk} = \underline{m}_{jk} = \arg\max_{m_{jk}} f(m_{jk})$

Proof. Our goal is to maximise $f(m_{jk}) = \frac{a}{2}m_{jk}^2 + bm_{jk} + \frac{c}{2}\log(m_{jk}^2)$. Take derivative with respect to m_{jk}

$$\frac{d}{dm_{jk}} = am_{jk} + b + c/m_{jk} = 0 \implies am_{jk}^2 + bm_{jk} + c = 0.$$

The roots are $\underline{\mathrm{m}}_{jk} := \frac{-b-\sqrt{b^2-4ac}}{2a}$ and $\overline{m}_{jk} := \frac{-b+\sqrt{b^2-4ac}}{2a}$.

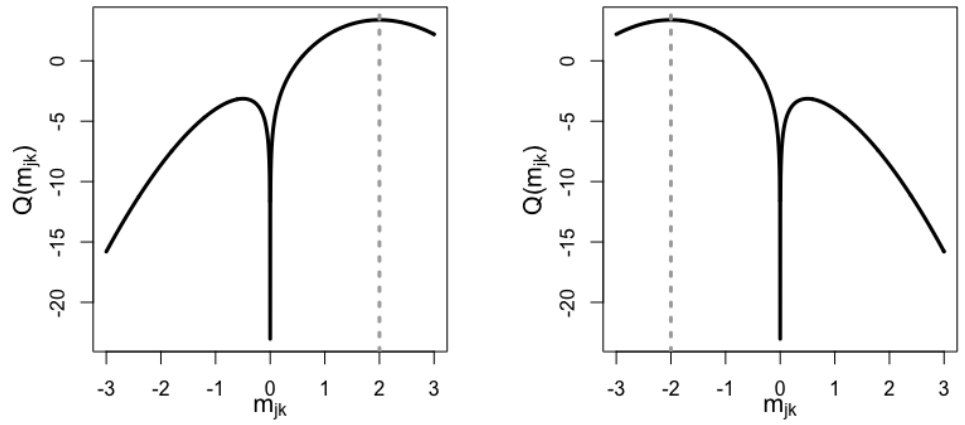
If $f(\overline{m}_{jk}) - f(\underline{\mathrm{m}}_{jk}) > 0$ then the global max is $\overline{m}_{jk}$, else the global max is $\underline{\mathrm{m}}_{jk}$. After

trivial algebra, $f(\overline{m}_{jk}) - f(\underline{m}_{jk}) = \frac{b}{2a}\sqrt{b^2-4ac} + c\log\left(\left[\frac{-b+\sqrt{b^2-4ac}}{b+\sqrt{b^2-4ac}}\right]^2\right)$.

For ease of notation let $z = \sqrt{b^2-4ac}$. Note that $z > 0$ and that, since $a < 0$, $c > 0$, that implies that $z - b > 0$. Then $f(\overline{m}_{jk}) - f(\underline{m}_{jk}) > 0$ if and only if $\frac{bz}{2a} > c\log\left(\left[\frac{z+b}{z-b}\right]^2\right) = 2c\log\left(\left[\frac{z+b}{z-b}\right]\right)$. Equivalently, $f(\overline{m}_{jk}) - f(\underline{m}_{jk}) > 0$ if and only if $\frac{bz}{4ac} > \log(z+b) - \log(z-b)$.

- Suppose $b > 0$. Then left-hand side is < 0, and right-hand side is > 0. Hence $f(\overline{m}_{jk}) - f(\underline{m}_{jk}) < 0 \implies$ global maximum is $\underline{m}_{jk}$
- Suppose $b < 0$. Then left-hand side is > 0, and right-hand side is < 0. Hence $f(\overline{m}_{jk}) - f(\underline{m}_{jk}) > 0 \implies$ global maximum is $\overline{m}_{jk}$ □

Figure 9 provides a visual representation of the global maximum for $\widehat{m}_{jk}$ under MOM-SS. Finally, the updates for $\widehat{\mathcal{T}}$ and $\widehat{\zeta}_k$ are equivalent to the ones obtained for Normal-SS.



(a) $b > 0$ (b) $b < 0$

Figure 9. Maximising m_{jk} for MOM-SS.

2.9 Laplace-Spike-and-MOM-Slab

2.9.1 Formulation

Another natural extension is to use Laplace-tailored NLPs based on the Spike-and-Slab LASSO in Equation (2.31). As illustrated in Figure 22 (right panels) Laplace-SS can help

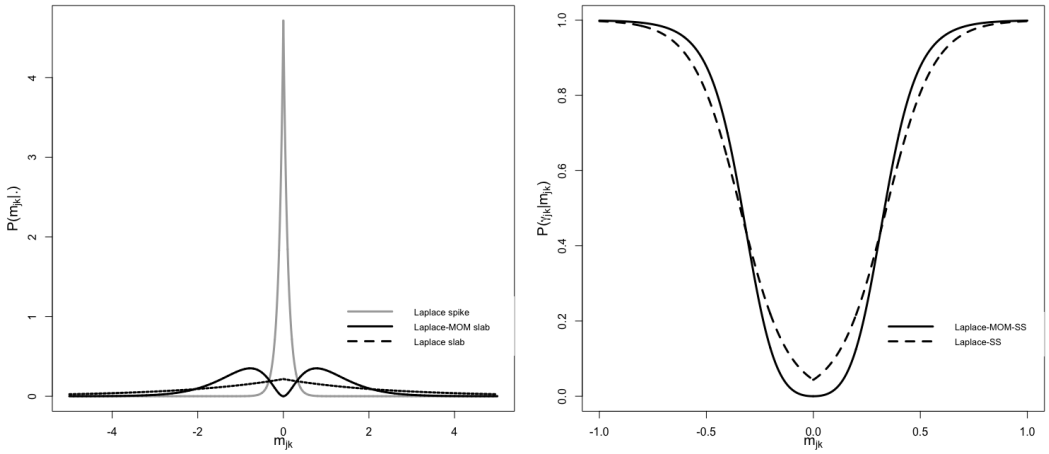


Figure 10. Prior comparison (left panel) for m_{jk} under Laplace-SS and Laplace-MOM-SS and its inclusion probabilities $\text{p}(\gamma_{jk} \mid m_{jk})$ (right panels). Scales (λ_0, λ_1) are set to the defaults from Section 2.10.

encourage sparsity, setting $\text{p}(\gamma_{jk} = 1 \mid m_{jk} = 0)$ to substantially smaller values (though still non-zero) than the Normal-SS.

Akin to (2.36), one could set a moment penalty on the Laplace density.

$$\begin{aligned} \text{p}(m_{jk} \mid \gamma_{jk} = 0, \widetilde{\lambda}_0) &= \text{Laplace}(m_{jk}; 0, \widetilde{\lambda}_0), \\ \text{p}(m_{jk} \mid \gamma_{jk} = 1, \widetilde{\lambda}_1) &= \frac{m_{jk}^2}{2\widetilde{\lambda}_1^2} \text{Laplace}(m_{jk}; 0, \widetilde{\lambda}_1). \end{aligned} \tag{2.41}$$

We denote (2.41) as Laplace-MOM-SS. Relative to (2.36), as illustrated in Figure 10, Laplace-MOM-SS leads to lower $\text{p}(\gamma_{jk} = 1 \mid m_{jk} = 0)$ and higher $\text{p}(\gamma_{jk} = 1 \mid m_{jk})$ for moderately large m_{jk}.

We derive an EM algorithm in Section 2.9.2 and discuss prior elicitation for Laplace-MOM-SS in Section 2.10 but in our examples we focus on the MOM-SS for simplicity. However, the Laplace-based (2.41) also leads to closed-form EM updates, as we now show.

2.9.2 EM algorithm for Laplace-MOM-SS

The E-step. For Laplace-MOM-SS the expected complete-data log-posterior can be split into $Q = C + Q_1 + Q_2$, with Q_2 as in (2.24) and

$$\begin{aligned} Q_1(\theta, M, \beta, \mathcal{T}) = &-\frac{1}{2}\sum_{i=1}^{n}\Big[\mathsf{x}_i^\top \mathcal{T}\mathsf{x}_i - 2\mathsf{x}_i^\top \mathcal{T} M \mathbb{E}[\mathsf{z}_i \mid \widehat{\Delta}, X] \\ &+ \operatorname{tr}\left(M^\top \mathcal{T} M \mathbb{E}[\mathsf{z}_i \mathsf{z}_i^\top \mid \widehat{\Delta}, X]\right)\Big] + \frac{n+\eta-2}{2}\log \mid \mathcal{T} \mid - \frac{\eta\xi}{2}\operatorname{tr}(\mathcal{T}) \\ &- \sum_{j=1}^{p}\sum_{k=1}^{q} \mid m_{jk} \mid \mathbb{E}\left[\frac{(1-\gamma_{jk})}{\widetilde{\lambda}_0} + \frac{(\gamma_{jk})}{\widetilde{\lambda}_1} \mid \widehat{\Delta}\right] \\ &+ \sum_{j=1}^{p}\sum_{k=1}^{q} \mathbb{E}[\gamma_{jk} \mid \widehat{\Delta}] \log(m_{jk}^2). \end{aligned} \tag{2.42}$$

The conditional expectation for the inclusion probability $\mathbb{E}[\gamma_{jk} \mid \widehat{\Delta}] = \widehat{p}_{jk}$ is

$$\widehat{p}_{jk} = \frac{1}{1 + \frac{2\widetilde{\lambda}_1^2}{\widehat{m}_{jk}^2}\frac{\widetilde{\lambda}_1}{\widetilde{\lambda}_0}\exp\left(- \mid \widehat{m}_{jk} \mid \left(\frac{1}{\widetilde{\lambda}_0} - \frac{1}{\widetilde{\lambda}_1}\right)\right)\frac{1-\mathbb{E}[\zeta_j]}{\mathbb{E}[\zeta_j]}} \tag{2.43}$$

and $\mathbb{E}\left[\frac{1}{(1-\gamma_{jk})\widetilde{\lambda}_0 + \gamma_{jk}\widetilde{\lambda}_1} \mid \widehat{\Delta}\right] = \frac{1-\widehat{p}_{jk}}{\widetilde{\lambda}_0} + \frac{\widehat{p}_{jk}}{\widetilde{\lambda}_1}$. $\mathbb{E}[\mathsf{z}_i \mid \widehat{\Delta}, X]$ and $\mathbb{E}[\mathsf{z}_i \mathsf{z}_i^\top \mid \widehat{\Delta}, X]$ are the same as in Equations (2.13) and (2.14) respectively.

The M-step: of the loadings, we consider using a coordinate descent algorithm (CDA) that performs successive univariate optimisation with respect to each m_{jk}. Notice that when $m_{jk} = 0$, the value of $Q(m_{jk} = 0) = -\infty$, thus the solution for the optimisation is given by setting $\frac{\partial Q_1}{\partial m_{jk}} = 0$.

The partial derivative of (2.42) w.r.t. M is

$$\begin{aligned} \frac{\partial Q}{\partial M} = &\sum_{i=1}^{n}\left[\widehat{\mathcal{T}}\mathsf{x}_i \mathbb{E}[\mathsf{z}_i^\top \mid \widehat{\Delta}, X] - \widehat{\mathcal{T}}\widehat{M}\mathbb{E}[\mathsf{z}_i \mathsf{z}_i^\top \mid \widehat{\Delta}, X]\right] \\ &- D^{\gamma, M} + 2\mathbb{E}[\gamma \mid \widehat{\Delta}] \circ \widehat{M}_{inv} = 0, \end{aligned} \tag{2.44}$$

with $D^{\gamma,M} \in \mathbb{R}^{p\times q}$ with element jk: $d_{jk}^{\gamma,M} = \operatorname{sign}(\widehat{m}_{jk})\mathbb{E}\left[\frac{1}{(1-\gamma_{jk})\widetilde{\lambda}_0 + \gamma_{jk}\widetilde{\lambda}_1} \mid \widehat{\Delta}\right]$, $\widehat{M}_{inv}$ a matrix with elements $1/\widehat{m}_{jk}$ and $A \circ B$ being the Hadamard (element-wise) product of two matrices A and B.

Taking the partial derivative of (2.42) with respect to m_{jk}, when $m_{jk} \neq 0$ and setting

it to 0 we obtain:

$$\begin{aligned}
\frac{\partial Q_1}{\partial m_{jk}} =& -\left(\sum_{i=1}^{n} \widehat{\tau}_{jj}\mathbb{E}[z_{ik}z_{ik}^{\top} \mid \widehat{\Delta}, X]\right)\widehat{m}_{jk} + \left(\sum_{i=1}^{n}\left[\widehat{\tau}_{jj}x_{ij}\mathbb{E}[z_{ik} \mid \widehat{\Delta}, X]\right.\right. \\
&\left.\left. - \sum_{r\neq k}^{q} \widehat{m}_{jr}\widehat{\tau}jj\mathbb{E}[z_{ir}z_{ik}^{\top} \mid \widehat{\Delta}, X]\right] - \operatorname{sign}(\widehat{m}_{jk})\mathbb{E}\left[\frac{1}{(1-\gamma_{jk})\widetilde{\lambda_0} + \gamma_{jk}\widetilde{\lambda_1}} \mid \widehat{\Delta}\right]\right) + \frac{2\mathbb{E}[\gamma_{jk}]}{\widehat{m}_{jk}} = 0 \\
\Leftrightarrow& -\left(\sum_{i=1}^{n} \widehat{\tau}_{jj}\mathbb{E}[z_{ik}z_{ik}^{\top} \mid \widehat{\Delta}, X]\right)\widehat{m}_{jk}^2 + \left(\sum_{i=1}^{n}\left[\widehat{\tau}_{jj}x_{ij}\mathbb{E}[z_{ik} \mid \widehat{\Delta}, X]\right.\right. \\
&\left.\left. - \sum_{r\neq k}^{q} \widehat{m}_{jr}\widehat{\tau}_j^{\top}\mathsf{b}_i\mathbb{E}[z_{ir}z_{ik}^{\top} \mid \widehat{\Delta}, X]\right] - \operatorname{sign}(\widehat{m}_{jk})\left[\frac{1-\widehat{p}_{jk}}{\widetilde{\lambda_0}} + \frac{\widehat{p}_{jk}}{\widetilde{\lambda_1}}\right]\right)\widehat{m}_{jk} + 2\mathbb{E}[\gamma_{jk}] \\
=& a\widehat{m}_{jk}^2 + b\widehat{m}_{jk} + c\cdot\operatorname{sign}(\widehat{m}_{jk})\widehat{m}_{jk} + d = 0
\end{aligned} \tag{2.45}$$

for $j = 1, \ldots, p$. We emphasise that when $m_{jk} = 0$, the only point of non-differentiability, $Q_1(m_{jk} = 0) = -\infty$. Thus the solution for m_{jk} is given by setting $\frac{\partial Q_1}{\partial m_{jk}} = 0$ as given in Lemma 2.4.

Lemma 2.4. *Let* $f(m_{jk}) = \frac{a}{2}m_{jk}^2 + bm_{jk} + c|m_{jk}| + \frac{d}{2}\log(m_{jk}^2)$, *where* $a < 0$, $c < 0$ *and* $d > 0$. *Define* $m_{jk}^{+} = \frac{-(b+c)-\sqrt{(b+c)^2-4ad}}{2a}$ *and* $m_{jk}^{-} = \frac{-(b-c)+\sqrt{(b-c)^2-4ad}}{2a}$.

If $b > 0$, *then* $m_{jk}^{+} = \arg\max_{m_{jk}} f(m_{jk})$. *If* $b < 0$, *then* $m_{jk}^{-} = \arg\max_{m_{jk}} f(m_{jk})$. *If* $b = 0$, *then* $m_{jk}^{+} = m_{jk}^{-} = \arg\max_{m_{jk}} f(m_{jk})$.

Proof. We aim to find the maximum of $f(m_{jk}) = \frac{a}{2}m_{jk}^2 + bm_{jk} + c|m_{jk}| + \frac{d}{2}\log(m_{jk}^2)$, where $a < 0$, $c < 0$ and $d > 0$. Note that when $m_{jk} = 0$, $Q_1(m_{jk} = 0) = -\infty$. Thus, the maximum of f is one of its critical points. Setting $\frac{\partial Q_1}{\partial m_{jk}} = 0$, we obtain

$$\frac{\partial Q_1}{\partial m_{jk}} = am_{jk} + b + c\cdot\operatorname{sign}(m_{jk}) + d/m_{jk} = 0 \implies am_{jk}^2 + bm_{jk} + c\cdot\operatorname{sign}(m_{jk})m_{jk} + d = 0.$$

- For $m_{jk} > 0$, we look for the solutions of $am_{jk}^2 + (b+c)m_{jk} + d = 0$.

  The roots of this polynomial are $\frac{-(b+c)\pm\sqrt{(b+c)^2-4ad}}{2a}$. Note that $\sqrt{(b+c)^2 - 4ad} > |b+c|$ since $a < 0$ and $d > 0$. Hence, the only acceptable root is $m_{jk}^{+} := \frac{-(b+c)-\sqrt{(b+c)^2-4ad}}{2a} > 0$, as the other one is negative.

- For $m_{jk} < 0$, we look for the solutions of $am_{jk}^2 + (b-c)m_{jk} + d = 0$.

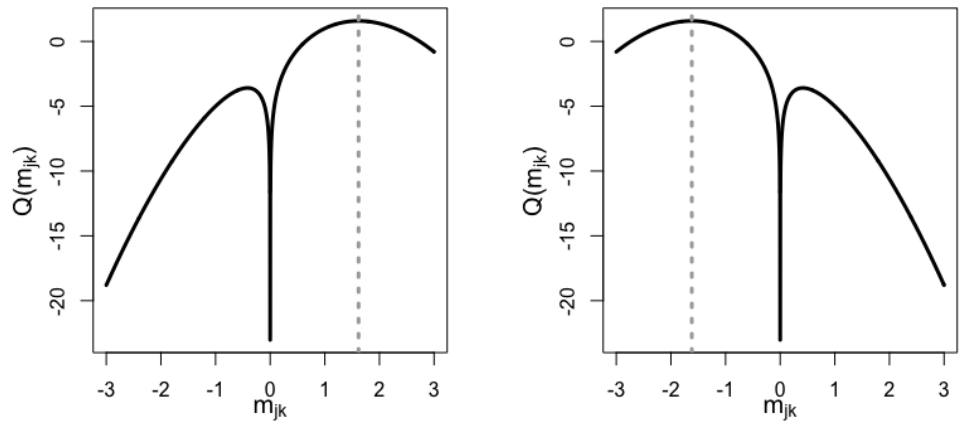
  The roots of this polynomial are $\frac{-(b-c)\pm\sqrt{(b-c)^2-4ad}}{2a}$. As before, $\sqrt{(b-c)^2 - 4ad} > |b-c|$. Hence, the only acceptable root is $m_{jk}^{-} := \frac{-(b-c)+\sqrt{(b-c)^2-4ad}}{2a} < 0$, as the

other one is positive.

- Suppose $b = 0$. Then clearly $f(m_{jk}) = f(-m_{jk})$ for all m_{jk}, i.e. the function is even. Therefore, m_{jk}^{+} and m_{jk}^{-} are opposite and both arg maxima.
- Suppose $b > 0$. By definition of f, $f(m_{jk}) > f(-m_{jk})$ for all $m_{jk} > 0$. In particular, $\max_{m_{jk}>0} f(m_{jk}) \geq \max_{m_{jk}<0} f(m_{jk})$ and $m_{jk}^{+} = \arg\max_{m_{jk}} f(m_{jk})$.
- Suppose $b < 0$. Then $f(m_{jk}) < f(-m_{jk})$ for all $m_{jk} > 0$. In particular, $\max_{m_{jk}>0} f(m_{jk}) \leq \max_{m_{jk}<0} f(m_{jk})$ and $m_{jk}^{-} = \arg\max_{m_{jk}} f(m_{jk})$. □

We remark that if x_i is continuous, the event of $b = 0$ has zero probability. If x_i is discrete and in presence of the rare event of $b = 0$, then the sign of the update for m_{jk} is set to the previous one. Figure 11 provides a visual representation of the update for $\widehat{m}_{jk}$ for Laplace-MOM-SS.

Finally, the updates for $\widehat{\mathcal{T}}$ and $\widehat{\zeta}_k$ are as in Equations (2.28) and (2.30).



(A) $b > 0$ (B) $b < 0$

Figure 11. Maximising m_{jk} for Laplace-MOM-SS.

2.10 Prior elicitation

A crucial aspect in a spike-and-slab prior is the choice of the prior scale parameters. It is common to fix the variance of the spike distribution λ_0 to a value close to zero. Regarding λ_1, one option is to set a hyper-prior or to try to estimate it from the data (George and

McCulloch, 1993, 1997; Ročková and George, 2014, 2018). Setting a hyper-prior does not bypass prior elicitation, as one then needs to set the hyper-prior parameters, whereas estimating λ_1 from the data increases the cost of computations. Instead, we capitalise on the fact that factor loadings have a natural interpretation in terms of the fraction of explained variance in X. Thus, we propose default values that dictate which coefficients are considered as meaningfully different from zero. These defaults are guidelines in the absence of a priori knowledge. A convenient feature of such an elicitation is that it can be easily extended to local priors and other non-Gaussian spike-and-slab priors.

Our goal is to find values $\widetilde{\lambda}_0$ and $\widetilde{\lambda}_1$ for the MOM-SS that distinguish practically relevant factors. In the absence of covariates, the factor model decomposes the total variance in variable j as $\text{Var}(x_{ij}) = \sum_{k=1}^{q} m_{jk}^2 + \tau_{jj}^{-1}$. Since we take the data to be standardised to unit variance, m_{jk}^2 can be roughly interpreted as the proportion of variance in variable j explained by factor k. We take $m_{jk}^2 > 0.1$ as a threshold for practical relevance. Specifically, we set $\widetilde{\lambda}_0$ such that $\text{p}(|m_{jk}| \leq \sqrt{0.1} \mid \widetilde{\lambda}_0) = 0.95$, that is $\widetilde{\lambda}_0 = \frac{0.1}{(\Phi^{-1}(0.025))^2} = 0.026$, where Φ^{-1} denotes the standard normal quantile function. Likewise we set $\text{p}(|m_{jk}| \geq \sqrt{0.1} \mid \widetilde{\lambda}_1) = 0.95$ under the MOM-SS, obtaining the default $\widetilde{\lambda}_1 = 0.2842$.

Regarding the Normal-SS prior,we set $\lambda_0 = \widetilde{\lambda}_0$ and λ_1 such that it is comparable to the MOM-SS in terms of informativeness, namely it matches the variance of the MOM-SS, obtaining that $\lambda_1 = 3\widetilde{\lambda}_1 = 0.8526$.

In the Laplace-MOM-SS, we analogously set $\widetilde{\lambda}_0 = -\frac{\sqrt{0.1}}{\log(0.05)} = 0.1056$ so that $\text{p}(|m_{jk}| \leq \sqrt{0.1} \mid \widetilde{\lambda}_0)$ and $\widetilde{\lambda}_1 = 0.3867$ such that $\text{p}(|m_{jk}| \geq \sqrt{0.1} \mid \widetilde{\lambda}_1) = 0.95$ for the Laplace-spike-and-MOM-slab prior. Finally for the Laplace-SS we set $\lambda_1 = 6\widetilde{\lambda}_1 = 0.9473$ and $\lambda_0 = \widetilde{\lambda}_0$ for the spike and slab component, respectively, matching the variances of the non-local Laplace-based priors.

The resulting priors are in Figures 8 and 10. We remark that going from Normal-SS to Laplace-SS has a big effect, particularly in the conditional inclusion probability around $m_{jk} = 0$, however, this effect is less marked for the Normal MOM-SS versus Laplace-MOM-SS. For this reason we will focus on the Normal MOM-SS in our examples. Deeper analysis of Laplace-based non-local priors would be interesting future work.

2.11 Initialisation of parameters

The EM algorithm has proved to be an efficient inference tool, it increases the log-posterior at each iteration by updating Δ at each iteration, however it does not guarantee global convergence and it can be sensitive to parameter initialisation. Poor initialisations can lead to slow convergence times, see e.g. Ročková and George (2017) for a discussion. We propose two different strategies: principal component and principal component with

rotation.

The first option is starting the algorithm with the Principal component solution for Factor analysis in Section 2.2.1. More formally, consider the eigendecomposition of the sample covariance matrix $\frac{1}{n}X^\top X$ where $l_1 \geq l_2 \geq l_q$ are the eigenvalues and $u_1, \ldots, u_q$ the eigenvectors. Set

$$M^{(0)} = [\sqrt{l_1}u_1 \mid \cdots \mid \sqrt{l_q}u_q]$$

and

$$\tau_{jj}^{-1(0)} = \max\{0, S_{jj} - \widehat{m}_j^\top \widehat{m}_j\},$$

where S_{jj} is the (j, j) element of $S = \frac{1}{n}X^\top X$.

The rotated principal components adds an extra step, a varimax rotation for the loadings obtained in the previous option. The reason for this extra step is to help escape local modes. The EM algorithm does not guarantee convergence to a global maximum, but it increases the log-posterior at each iteration. This local maxima issue is intensified by the non-identifiability of the factor model through the rotational ambiguity of the likelihood and the strong association between the updates of loadings and factors.

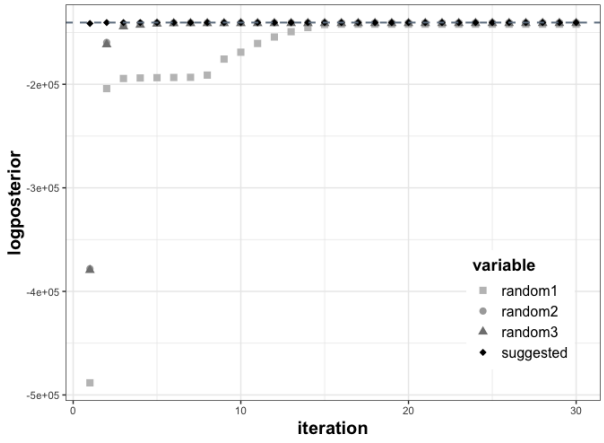


Figure 12. Comparison of the log-posterior convergence at four different initialisations of parameters. The suggested initialisation in squared shape and the log-posterior evaluated at the data-generating truth in dotted line.

2.12 Post-processing for model selection and dimensionality reduction

The EM algorithm gives point estimates $(\widehat{M}, \widehat{\mathcal{T}}, \widehat{\zeta})$. Under Laplace-SS one can obtain exact sparsity via $\widehat{m}_{jk} = 0$, however this is not the case for our other priors. To address

this, we define $\widehat{\gamma}$ as the solution of the following optimisation problem

$$\widehat{\gamma} = \text{argmax}_{\gamma} \text{p}(\gamma \mid X, \widehat{M}, \widehat{\mathcal{T}}, \widehat{\zeta}) = \text{argmax}_{\gamma} \prod_{jk} \text{p}(\gamma_{jk} | \widehat{m}_{jk}, \widehat{\zeta}_k) \tag{2.46}$$

where the right-hand side follows from the assumed conditional independence of m_{jk}. That is, we set $\widehat{\gamma}_{jk} = 1$ if $\text{p}(\gamma_{jk} = 1 | \widehat{m}_{jk}, \widehat{\zeta}_k) > 0.5$ and $\gamma_{jk} = 0$ otherwise. When $\widehat{\gamma}_{jk} = 0$ we set $\widehat{m}_{jk} = 0$ effectively selecting the number of factors and the non-zero loadings within each factor.

As an alternative post-processing step we consider that in some applications one may want to select only the number of factors. We then consider to setting $\widetilde{\gamma}_{jk} = 1$ if $\sum_{j=1}^{p} \widehat{\gamma}_{jk} \neq 0$ and $\widetilde{\gamma}_{jk} = 0$ otherwise.

The combination of the two initialisation alternatives and two different post-processing options gives four possible solutions for $\widehat{M}$. To choose which is best in our examples, we use weighted 10-fold cross-validation, selecting the model with smallest weighted cross validation reconstruction error. Algorithm 3 provides a pseudo-code-algorithm for the weighted 10-fold cross-validation used through this thesis.

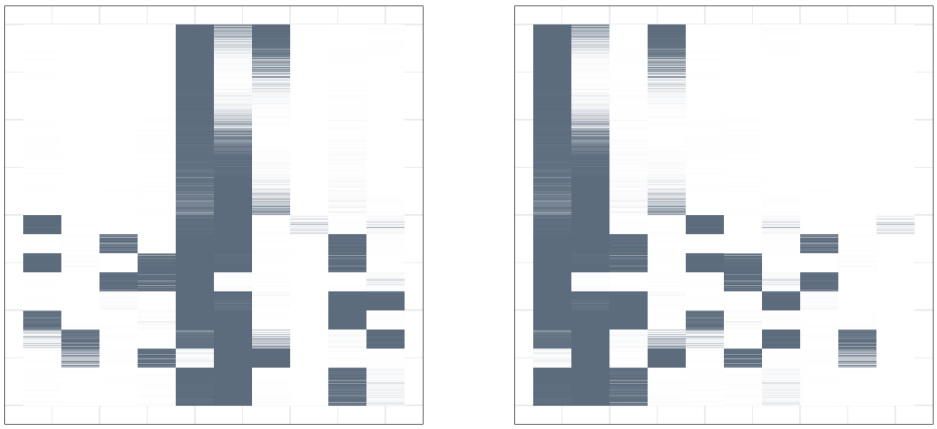
Algorithm 3: Weighted 10-fold cross-validation for Bayesian factor analysis

initialise $\varepsilon_X = 0$
set 10 random cross-validation subsets of
  Observations: $\{\mathsf{x}^{[1]}, \ldots, \mathsf{x}^{[10]}\} \in \mathbb{R}^{\frac{n}{10} \times p}$
for $r \leftarrow 1, \ldots, 10$ **do**
  set: Cross-validation subsets
  $\widetilde{\mathsf{x}} := (\mathsf{x}^{[1]}, \ldots, \mathsf{x}^{[r-1]}, \mathsf{x}^{[r+1]}, \ldots, \mathsf{x}^{[10]})$
  compute: EM algorithm
    input: $\widetilde{\mathsf{x}}$
    output: $\widehat{M}, \widehat{\mathcal{T}}, \widehat{\theta}, \widehat{\beta}, \widehat{\zeta}$
  set: Test factors $\widehat{\mathsf{z}}_i = (\mathbf{I}_q + \widehat{M}^\top \widehat{\mathcal{T}} \widehat{M})^{-1} \widehat{M}^\top \widehat{\mathcal{T}} \mathsf{x}^{[r]}$
  compute $\varepsilon_X = \varepsilon_X + \sum_i || \left[\mathsf{x}_i^{[r]} - \widehat{M} \widehat{\mathsf{z}}_i \right] \widehat{\mathcal{T}} ||_F$
end
set $\varepsilon_X = \frac{\varepsilon_X}{10}$

Finally we re-order of the factors so that $\sum_{j=1}^{p} \gamma_{jk}$ is decreasing in k, which under our prior (2.20) is guaranteed to increase the log-posterior. This is the so-called left-ordered inclusion matrix of Griffiths and Ghahramani (2011) and facilitates the interpretation of latent factors.

Definition 2.5. Let $\|\gamma_{\cdot k}\|_0 = \sum_{j=1}^{p} \gamma_{jk}$ be the L0 norm of $\gamma_{\cdot k}$, i.e. the number of non-zero entries of the column k of matrix γ. The left-ordered function lof orders the columns of

a binary matrix from left to right according to their magnitude. Namely, lof permutes the columns of γ in such a way that $\|\gamma_{\cdot 1}\|_0 \geq \cdots \geq \|\gamma_{\cdot k}\|_q$, obtaining a new ordered $\gamma^* = \text{lof}(\gamma)$ (see Figure 13). In case two or more columns have the same magnitude, their order is not changed by convention.



(A) Non-left-ordered γ (B) Left-ordered γ

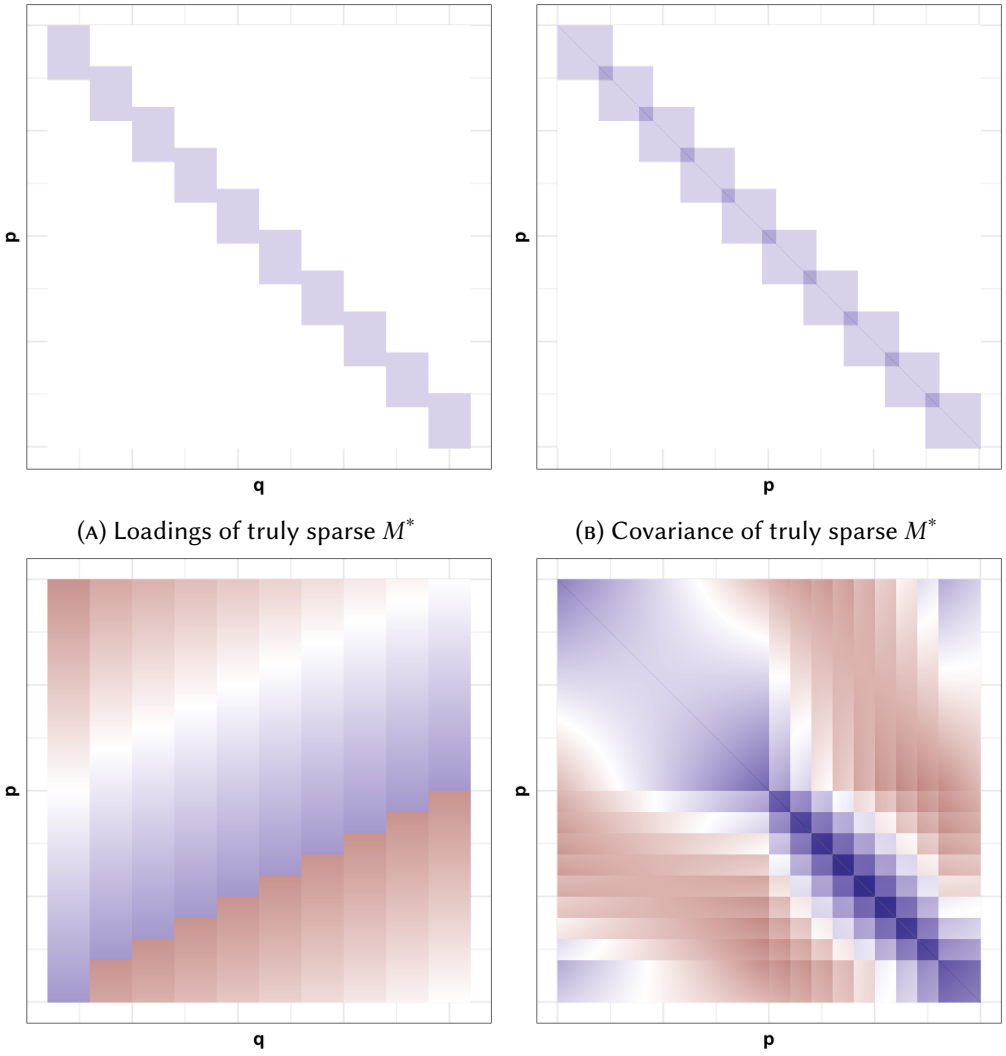
Figure 13. The γ matrix on the left is transformed into the left-ordered binary matrix on the right γ^* by the function $\text{lof}(\gamma) = \gamma^*$. Zero entries are in white, non-zero entries in gray.

2.13 Simulation studies

We first assess our approach on simulated data. We study simulations under two different loading matrices M (truly sparse and dense). We compare our methods with the Fast Bayesian Factor Analysis via Automatic Rotations to Sparsity (FastBFA) of Ročková and George (2017) and the Penalized Likelihood Factor Analysis with a LASSO penalty (LASSO-BIC) of Hirose and Yamamoto (2015).

Our **R** package is available at `https://github.com/AleAviP/BFR.BE` (See Appendix A). We used **R** function `FACTOR_ROTATE` of Ročková and George (2017) for FastBFA, the **R** package `fanc 2.2` for LASSO-BIC (Hirose et al., 2016).

Prior parameters for the Normal-SS and MOM-SS were set as in Section 2.10, the hyper-parameters for FastBFA were set via Dynamic Posterior Exploration as in Ročková and George (2017) with $1/\lambda_0 = 0.001$ and $1/\lambda_1 \in \{5, 10, 20\}$ and using varimax robustifications. For the LASSO-BIC we selected the model with smallest BIC to set the regularization parameter.



(A) Loadings of truly sparse M^*

(B) Covariance of truly sparse M^*

(C) Loadings of dense M^*

(D) Covariance of dense M^*

Figure 14. Synthetic data. Heatmaps of data-generating loadings and covariance with red highly negative, blue highly positive and white zero values.

To assess the precision of the parameter estimates returned by the EM algorithm, we simulated data from two different data-generating truths: truly sparse and dense loadings M. In both, the truth was set to $q^* = 10$ factors. The dense loadings matrix has a grid of elements set uniformly between $(-1, 1)$, whereas the truly sparse M has a banded-diagonal structure with $m_{jk} = 1$ for the non-zero elements, as shown in Figure 14.

We simulated $n = 100$ observations from $\mathsf{x}_i = M^*\mathsf{z}_i, +\mathsf{e}_i$, with growing $p = 1,000$ and $1,500$, where the factors $\mathsf{z}_i \sim N(0, \mathbf{I}_q)$, the errors $\mathsf{e}_i \sim N(0, \mathcal{T}^{-1})$ with $\mathcal{T}^{-1} = \mathbf{I}_p$, and the loadings M^* are set as dense or sparse as in Figure 14. For comparison, FastBFA was

initialised as our models via principal components (Section 2.11). Tables 1 and 2 show the selected number of factors $\widehat{q}$, the number of estimated non-zero loadings $||\widehat{M}||_0 = \sum_{j,k} \mathbb{1}(\widehat{m}_{jk} \neq 0)$, the Frobenius norm (F.N.) between the true expected value and its reconstruction $||E[X] - \widehat{E}[X]||_F = ||ZM^\top - \mathbb{E}[Z \mid \widehat{\Delta}, X]\widehat{M}^\top||_F$ and between the true and reconstructed covariances $||\text{Cov}[x_i] - \widehat{\text{Cov}}[x_i]||_F = ||(MM^\top + \mathcal{T}^{-1}) - (\widehat{M}\widehat{M}^\top + \widehat{\mathcal{T}}^{-1})||_F$, and the number of iterations until convergence. The mean across 100 independent simulations is displayed and the model with smallest mean Frobenius norm per scenario is indicated in bold.

2.13.1 Dense loadings

Table 1: Synthetic data without batch effects for $n = 100$, $q^* = 10$, $p = 1,000$ or $1,500$ parameters, dense loadings M^*.

	$p = 1,000$					$p = 1,500$				
Model	$\widehat{q}$	$\|\|\widehat{M}\|\|_0$	$\|\|\mathbb{E}[X] - \widehat{\mathbb{E}}[X]\|\|_F$	$\|\|\text{Cov}[x_i] - \widehat{\text{Cov}}[x_i]\|\|_F$	it	$\widehat{q}$	$\|\|\widehat{M}\|\|_0$	$\|\|\mathbb{E}[X] - \widehat{\mathbb{E}}[X]\|\|_F$	$\|\|\text{Cov}[x_i] - \widehat{\text{Cov}}[x_i]\|\|_F$	it
	$q = 10$									
Flat	10.0	10000.0	104.8	**1173.3**	2.0	10.0	15000.0	126.5	**1895.7**	2.0
Normal-SS	10.0	1859.7	**92.4**	1266.4	9.3	10.0	2461.0	**112.5**	1988.2	6.9
MOM-SS	10.0	1468.6	93.5	1294.3	9.7	10.0	2059.1	114.3	1998.5	6.3
FastBFA	9.6	976.9	137.9	1738.2	153.6	9.4	1400.4	163.2	5638.7	162.0
LASSO-BIC	10.0	5331.3	110.6	1682.5	NA	10.0	8607.7	137.4	2524.8	NA
	$q = 100$									
Flat	100.0	100000.0	313.2	1200.6	3.0	100.0	150000.0	376.2	1925.3	2.5
Normal-SS	34.8	3418.5	190.5	1190.7	4.2	14.8	5083.8	154.4	1911.8	4.0
MOM-SS	10.5	3215.9	108.9	1178.7	5.0	11.2	4232.8	135.6	1902.8	4.0
FastBFA	96.6	3379.3	297.4	**451.2**	11.3	97.3	4558.4	362.1	**670.5**	10.5
LASSO-BIC	11.0	4829.2	**80.5**	1682.7	NA	11.1	7839.6	**99.5**	2524.8	NA

Real-life datasets can contain various types of sparsity. However, we first considered the scenario where M is dense and one has guessed correctly the true number of factors $q = q^* = 10$. The aim of this setting was to investigate if MOM-SS shrinkage provided a poor estimation when the factors are not truly sparse. As shown in Table 1, MOM-SS and Normal-SS performed similarly as p grew, and competitively relative to the flat prior.

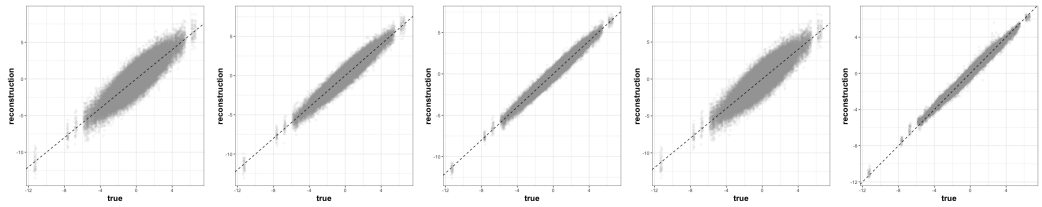
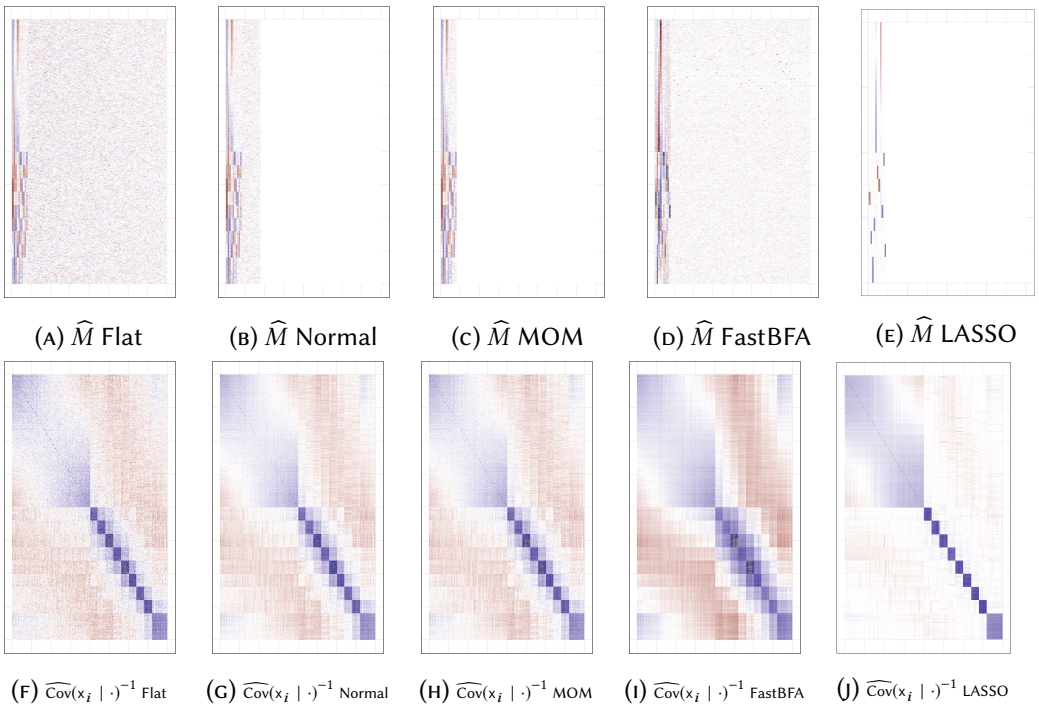


Figure 15. Scatterplots comparing $ZM^\top$ vs. $\mathbb{E}[Z \mid \widehat{\Delta}, X]\widehat{M}^\top$ between the different models under dense loadings M with $q = 100$ in simulations without batch effect.

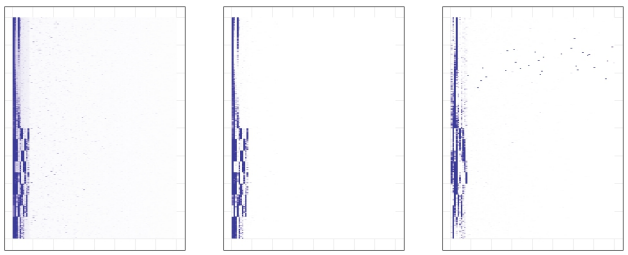
To extend our example, we then set $q = 100$ to illustrate the performance when there is sparsity in terms of the number of factors, but not within factors. LASSO-BIC had the best reconstruction for the mean but performed poorly on the covariance, whereas FastBFA outperformed all the models to estimate the covariance but performed poorly for the mean. However, MOM-SS had a good balance in terms of estimating the expected value and the covariance, being the second best in both cases.



(A) $\widehat{M}$ Flat (B) $\widehat{M}$ Normal (C) $\widehat{M}$ MOM (D) $\widehat{M}$ FastBFA (E) $\widehat{M}$ LASSO

(F) $\widehat{\text{Cov}}(x_i \mid \cdot)^{-1}$ Flat (G) $\widehat{\text{Cov}}(x_i \mid \cdot)^{-1}$ Normal (H) $\widehat{\text{Cov}}(x_i \mid \cdot)^{-1}$ MOM (I) $\widehat{\text{Cov}}(x_i \mid \cdot)^{-1}$ FastBFA (J) $\widehat{\text{Cov}}(x_i \mid \cdot)^{-1}$ LASSO

Figure 16. Heatmaps of loadings and covariance (red denotes large negative values, blue large positive values, white denotes zero).



(A) $\widehat{\gamma}$ (Normal) (B) $\widehat{\gamma}$ (MOM) (C) $\widehat{\gamma}$ (FastBFA)

Figure 17. Heatmaps of inclusion probability (white denotes 0, dark blue denotes 1).

Figure 15 presents a visual comparison of $ZM^\top$ vs. $\mathbb{E}[Z \mid \widehat{\Delta}, X]\widehat{M}^\top$ of this scenario.

Figures 16 and 17 display a visual representation of the estimated $\widehat{M}$, $\widehat{\text{Cov}}(\text{x}_i \mid \cdot)^{-1}$ and $\widehat{\gamma}$. It is important to notice that although FastBFA had the highest estimated cardinality $\widehat{q}$, the estimated $\widehat{\gamma}$ as very few values closer to zero after the 10th loading as seen in Figure 17.

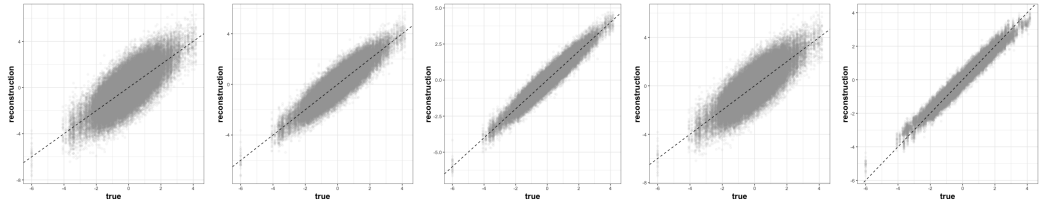
2.13.2 Truly sparse loadings

Table 2: Synthetic data without batch effects for $n = 100$, $q^* = 10$, $p = 1,000$ or $1,500$ parameters, truly sparse loadings M^*.

	$p = 1,000$					$p = 1,500$				
Model	$\widehat{q}$	$\|\|\widehat{M}\|\|_0$	$\|\|\mathbb{E}[X] - \widehat{\mathbb{E}}[X]\|\|_F$	$\|\|\text{Cov}[x_i] - \widehat{\text{Cov}}[x_i]\|\|_F$	it	$\widehat{q}$	$\|\|\widehat{M}\|\|_0$	$\|\|\mathbb{E}[X] - \widehat{\mathbb{E}}[X]\|\|_F$	$\|\|\text{Cov}[x_i] - \widehat{\text{Cov}}[x_i]\|\|_F$	it
	$q = 10$									
Flat	10.0	10000.0	104.8	184.1	2.0	10.0	15000.0	126.4	301.4	2.0
Normal-SS	10.0	1300.1	55.8	124.1	3.9	10.0	1942.1	**68.0**	248.5	3.0
MOM-SS	10.0	1299.9	**53.8**	**122.5**	4.3	10.0	1943.0	69.4	**235.2**	2.3
FastBFA	8.7	1076.3	74.8	176.8	93.1	7.1	1320.4	84.7	344.2	122.7
LASSO-BIC	10.0	5304.3	77.4	424.0	NA	10.0	8397.0	93.3	636.3	NA
	$q = 100$									
Flat	100.0	100000.0	313.7	310.8	3.0	100.0	150000.0	375.4	446.7	2.5
Normal-SS	22.0	2801.8	165.3	203.7	4.0	42.5	2795.9	230.0	335.1	4.3
MOM-SS	10.5	2156.8	109.7	**194.1**	5.0	11.2	2430.5	136.4	**324.8**	4.0
FastBFA	97.9	1508.9	283.0	215.2	9.9	97.6	2229.7	363.0	326.4	9.2
LASSO-BIC	10.0	4815.5	**75.0**	425.1	NA	10.0	7980.8	**91.2**	637.1	NA

We further illustrate our model under the arguably more interesting case of truly sparse loadings. First we set $q = 10$, the true cardinality. The results are in Table 2. In this scenario MOM-SS and Normal-SS presented the best results both to estimate $\mathbb{E}[X]$ and $\text{Cov}[x_i]$. This example reflects the advantages of shrinkage and the varimax rotation for the initialisation in the loadings, leading to good sparse solutions.

Finally we considered the same scenario with $q = 100$. LASSO-BIC was best to estimate the mean at the cost of reduced precision in the covariance reconstruction. MOM-SS displayed the lowest error for the covariance and second smallest for the mean, showing a good balance between those metrics.

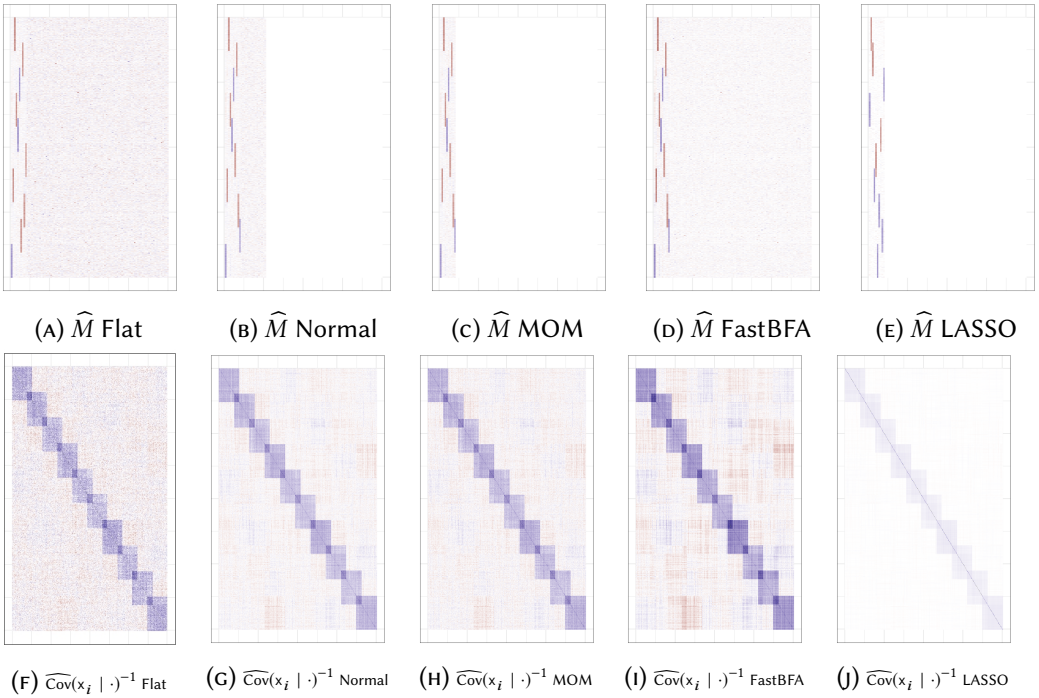


Flat Sparse M | Normal-SS Sparse M | MOM-SS Sparse M | FastBFA Sparse M | LASSO Sparse M

Figure 18. Scatterplots comparing $ZM^\top$ vs. $\mathbb{E}[Z \mid \widehat{\Delta}, X]\widehat{M}^\top$ between the different models under truly sparse loadings M with $q = 100$ in simulations without batch effect.

Figure 18 presents the scatterplots comparing $ZM^\top$ vs. $\mathbb{E}[Z \mid \widehat{\Delta}, X]\widehat{M}^\top$ and the heatmaps

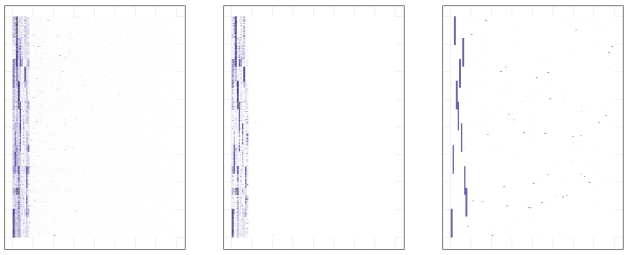
of the reconstruction of $\widehat{M}$, $\widehat{\mathrm{Cov}}(x_i \mid \cdot)^{-1}$ and $\widehat{\gamma}$ for the different models.



(A) $\widehat{M}$ Flat (B) $\widehat{M}$ Normal (C) $\widehat{M}$ MOM (D) $\widehat{M}$ FastBFA (E) $\widehat{M}$ LASSO

(F) $\widehat{\mathrm{Cov}}(x_i \mid \cdot)^{-1}$ Flat (G) $\widehat{\mathrm{Cov}}(x_i \mid \cdot)^{-1}$ Normal (H) $\widehat{\mathrm{Cov}}(x_i \mid \cdot)^{-1}$ MOM (I) $\widehat{\mathrm{Cov}}(x_i \mid \cdot)^{-1}$ FastBFA (J) $\widehat{\mathrm{Cov}}(x_i \mid \cdot)^{-1}$ LASSO

Figure 19. Heatmaps of loadings and covariance (red denotes large negative values, blue large positive values, white denotes zero).



(A) $\widehat{\gamma}$ (Normal) (B) $\widehat{\gamma}$ (MOM) (C) $\widehat{\gamma}$ (FastBFA)

Figure 20. Heatmaps of inclusion probability (white denotes 0, dark blue denotes 1).

Recall that we used a coordinate descent algorithm for the non-local prior, which as a potential drawback could require a larger number of iterations than jointly optimising multiple elements in M. However, Tables 1 and 2 showed that MOM-SS required roughly the same number of iterations to converge as the Normal-SS. We can see that MOM-SS and LASSO-BIC estimated $\widehat{q}$ accurately. Note that in general FastBFA had the highest estimated latent cardinality $\widehat{q}$, due to adding some columns of M that contain very few

non-zero loadings after the tenth factor, as shown in Figure 20. Nonetheless, this model displayed a mean number of non-zero loadings closer to the ground truth (1,300 and 1,940 for the $p = 1,000$ and $p = 1,500$ respectively). Overall, MOM-SS showed a competitive behaviour that was robust to changing the denseness of M, the number of variables p and maximum number of loadings.

2.14 Conclusions

In this Chapter we presented novel non-local spike-and-slab priors. To our knowledge this is the first time NLPs have been implemented in the factor analysis context. We gave deterministic optimisations for our model and provided novel EM algorithms to obtain closed-form posterior modes.

We showed that the use of sparse models increases the quality of our estimations. MOM-SS priors prove to be appealing, improving the estimation of factor cardinality and encouraging parsimony and selective shrinkage. In general, MOM-SS achieved a good balance between estimating the mean, which is useful for dimensionality reduction, and sparse covariance estimation.

3

Batch effect correction using Bayesian factor regression

3.1 Introduction

This chapter presents a novel model for joint unsupervised analysis of high-throughput data from different batches, i.e. generated under different experimental conditions, when new samples are incrementally added to existing samples, or in analyses coming from different projects, laboratories, or platforms.

We address dimensionality reduction via a model-based framework relying on Bayesian factor analysis and latent factor regression. Our model aims to account for systematic biases or sources of variation that do not reflect any underlying patterns of interest, i.e. batch effects. We build on the Model from Section 2. Strategies for batch effect correction include data pre-processing, for example via the so-called ComBat empirical Bayes approach (Johnson et al., 2007) or via singular value decomposition (SVD) (Leek and Storey, 2007). Another possibility is to use factor models to learn on the one hand biological patterns via common factors shared across the different data sources and on the other hand non-common sources of variation via data-specific factors (De Vito et al., 2018b,a). Those approaches while useful, do not provide a model-based approach for dimensionality reduction that corrects both mean and variance batch effects, and returns sparse loadings. We extend directly Section 2 by modelling observations with a regression on latent factors, observed covariates and batch effects that can alter the mean and intrinsic variance structures. Model fitting is done via an Expectation-Maximisation (EM) algorithm to obtain maximum posterior mode parameter estimates in a computationally efficient manner. We focus on three different continous prior formulations for the loadings discussed in Section 2: flat, Normal-spike-and-slab and our Normal-spike-and-MOM-slab. We also discuss non-local Laplace-tailed extensions. We obtain closed-form EM updates, a novel

contribution to the non-local prior literature.

The outline of this chapter is as follows. Section 3.2 reviews latent factor regression and introduces our extension, which includes a variance batch effect adjustment. Section 3.3 proposes prior formulations including non-local priors on the loadings. Section 3.4 describes several EM algorithms for model fitting, parameter initialisation and post-processing steps required for effective model selection and dimension reduction. Section 3.5 presents applications on simulations. Section 3.6 concludes. Software implementing our methodology is available at `https://github.com/AleAviP/BFR.BE`.

3.2 Latent factor regression with batch effects

Consider vectors $\mathsf{x}_i = (x_{i1}, x_{i2}, \ldots, x_{ip})^\top \in \mathbb{R}^p$, observed for $i = 1, \ldots, n$ individuals. The factor regression model defines x_i as a regression on p_v observed covariates denoted by $\mathsf{v}_i \in \mathbb{R}^{p_v}$, and q low-dimensional latent variables denoted $\mathsf{z}_i \in \mathbb{R}^q$, also known as latent coordinates or factors. The standard factor regression model is

$$\mathsf{x}_i = \theta \mathsf{v}_i + M\mathsf{z}_i + \mathsf{e}_i, \tag{3.1}$$

where $\theta \in \mathbb{R}^{p \times p_v}$ is the matrix of regression coefficients, and M, z_i and e_i are as in the standard FA model (2.1) .

Equation (3.1) regresses the observed data X on known covariates and on a latent factor structure. In particular, it allows additive batch effects to be accounted for by incorporating the variables recording the batches into v_i. However, in practice one often observes more complex batch effects; for example in bioinformatics it is common to observe multiplicative effects on the variance (Johnson et al., 2007). We will later describe an example of this, shown in Figure 30. Such artefacts cannot be captured by (3.1) given that Σ is assumed constant across all individuals.

To address this issue we extend (3.1) by allowing Σ to depend on i. Suppose the data were obtained in p_b batches, e.g. from different days, laboratories or instrumental calibrations, with n_l individuals in batch l, for $l = 1, \ldots, p_b$, such that $n_1 + n_2 + \cdots + n_{p_b} = n$. Let b_i be the indicator vector of length p_b defined as $b_{il} := 1$ if individual i is in batch l, $b_{il} := 0$ otherwise.

We incorporate batch effects by adding a mean and variance adjustment. We let

$$\mathsf{x}_i = \theta \mathsf{v}_i + M\mathsf{z}_i + \beta \mathsf{b}_i + \mathsf{e}_i, \tag{3.2}$$

where θ, v_i, M and z_i are as (3.1), $\beta \in \mathbb{R}^{p \times p_b}$ captures additive batch effects and the

variance of e_i captures multiplicative batch effects. We denote by τ_{jl}, $j = 1, \ldots, p$ and $l = 1, \ldots, p_b$ as the j^{th} idiosyncratic precision element in batch l. Then, given $b_{il} = 1$, the errors are independently distributed as $\mathsf{e}_{ij} \sim N(0, \tau_{jl}^{-1})$. Further, denote by $\mathcal{T}$ the $p \times p_b$ matrix that has τ_{jl} as its (j, l) element.

To help interpret the practical implications of the model, suppose that one has orthonormal factor loadings $M^\top M = \mathbf{I}$. Then (3.2) implies

$$\mathsf{z}_i = M^\top (\mathsf{x}_i - (\theta \mathsf{v}_i + \beta \mathsf{b}_i + \mathsf{e}_i)) \tag{3.3}$$

and thus, $\mathbb{E}(\mathsf{z}_i \mid \mathsf{x}_i, \mathsf{v}_i, \mathsf{b}_i, M, \theta, \beta) = M^\top \mathsf{x}_i - M^\top \theta \mathsf{v}_i - M^\top \beta \mathsf{b}_i$. That is, the mean of the latent coordinates is the projection $M^\top \mathsf{x}_i$ plus a translation given by the batch effect adjustment and (potentially) the observed covariates. An interesting observation is that their covariance $\mathrm{Cov}(\mathsf{z}_i \mid \mathsf{x}_i, \mathsf{v}_i, \mathsf{b}_i, M, \theta, \beta, \mathcal{T}) = M^\top \mathcal{T}_{\mathsf{b}_i}^{-1} M$ depends on the multiplicative batch-dependent noise. As an example, Figure 30(b) show the two first factors of an ovarian dataset pre-processed by ComBat. Relative to the unadjusted Figure 30(a), ComBat removes systematic differences in mean and variance accross the 2 batches, however the latent coordinates exhibit distinct covariances. To obtain suitably-adjusted low-dimension coordinates one should estimate $\mathcal{T}$ jointly with (M, θ, β).

Model (3.2) can be represented in matrix notation as

$$X = V\theta^\top + ZM^\top + B\beta^\top + E, \tag{3.4}$$

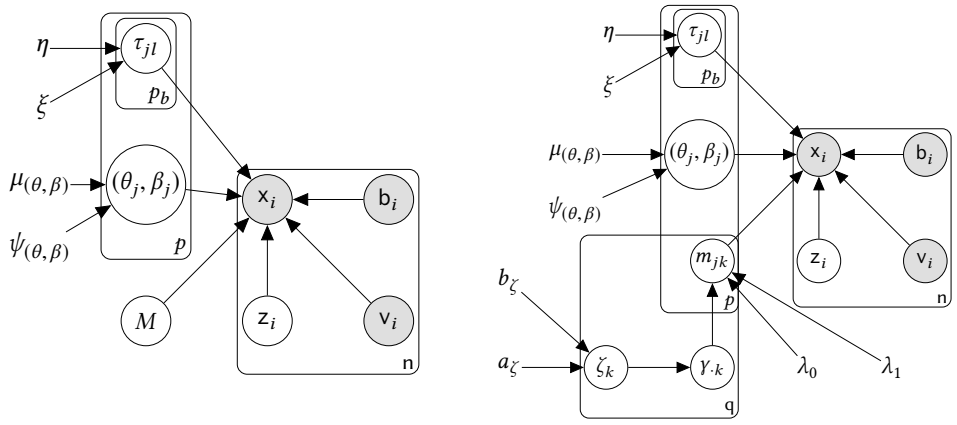
where $E \in \mathbb{R}^{n \times p}$ is the matrix of errors.

As mentioned in Section 2.2, the latent factor model is non-identifiable up to orthogonal transformations, of the form $M^{*\top} = A^\top M^\top$ and $Z^* = ZA$, where A is any orthogonal $q \times q$ matrix. Through this chapter we follow the same strategy as in the previous chapter, inducing sparse solutions via local and non-local penalties.

3.3 Prior formulation

To complete Model (3.2) we set priors for the loadings M, precisions τ_{jl}, and regression parameters (θ, β). Through our proposed default prior formulation we assume that the columns in X have been centred to zero mean and unit variance. For the idiosyncratic precisions τ_{jl} we set

$$\tau_{jl} \mid \eta, \xi \sim \mathrm{Gamma}(\eta/2, \eta\xi/2) \tag{3.5}$$



(A) Flat prior

(B) Spike-and-slab prior

FIGURE 21. Directed acyclic graph (DAG) for Bayesian factor regression with Batch Effect correction for different prior formulation: (a) Flat or non-sparse loading matrix. (b) Spike-and-slab or sparse loading matrix.

independently across $j = 1, \ldots, p$ and $l = 1, \ldots, p_b$. By default in our examples we set the fairly informative values $\eta = \xi = 1$, leading to diffuse though proper priors.

For the regression parameters we set

$$(\theta_j, \beta_j) \sim N(0, \psi \mathbf{I}), \quad j = 1, \ldots, p \tag{3.6}$$

where ψ is a user-defined prior dispersion that in our examples by default we set to $\psi = 1$. The choice of $\psi = 1$ assigns the same marginal prior variances to elements in (θ_j, β_j) as the unit information prior often adopted as a default for linear regression (Schwarz, 1978).

We remark that this prior does not encourage sparsity in the regression parameters (θ, β), which we view as reasonable provided the number of variables p_v and batches p_b are moderate. For large p_v or p_b, a direct extension of our prior on the loadings M could be adopted.

As shown in the Chapter 2, sparsity in the loadings matrix M plays a crucial role to ease interpretation. In this chapter we study the same five prior formulations:

- Flat: Equation (2.10);
- Normal-spike-and-slab (Normal-SS): Equation (2.21);
- Laplace-spike-and-slab (Laplace-SS): Equation (2.31);
- Normal-spike-and-MOM-slab (MOM-SS): Equation (2.36);
- Laplace-spike-and-MOM-slab (Laplace-MOM-SS): Equation (2.41).

We finish the prior specification with a hierarchical prior over the latent indicator $\gamma = \{\gamma_{jk}, j = 1, \ldots, p, k = 1, \ldots, q\}$ as in (2.20) (Chapter 2) for all the spike-and-slab priors.

Figure 21 provides the DAG for the Model 3.2 for Flat and spike-and-slab priors (SS) for the loadings.

3.4 Parameter estimation

Section 3.4.1 provides two EM algorithms to obtain posterior modes for our factor regression with batch effect correction with and without sparse formulation. At the core of these algorithms is the fact that, conditional on the data and all other model parameters, we can set $\widetilde{\mathsf{x}}_i = \mathsf{x}_i - \theta \mathsf{v}_i - \beta \mathsf{b}_i$ and express the model in (3.2) as a linear regression $\widetilde{\mathsf{x}}_i = M\mathsf{z}_i + \mathsf{e}_i$, where M and τ_{jl} are fixed at their current values of each maximisation step Section 3.4.2 outlines an algorithm separately for Normal-SS, MOM-SS, Laplace-SS and Laplace-MOM-SS priors. Section 3.4.3 discusses parameter initialisation and Section 3.4.4 how to post-process the fitted model to obtain sparse solutions and variance-adjusted dimensionality reduction.

3.4.1 EM algorithm under a uniform prior

We outline an EM algorithm to fit Model (3.2) under a uniform prior $\mathrm{p}(M) \propto 1$ on the loadings via maximum a posteriori (MAP) estimation. The algorithm maximises the log-posterior by treating the latent factors Z as missing data and setting them to their expectation (conditional on all other parameters) in the E-step. Then, the remaining parameters $\Delta = (M, \theta, \beta, \mathcal{T})$ are optimised in the M-step. In other words, the EM algorithm obtains a local mode of the log-posterior $\mathrm{p}(M, \theta, \beta, \mathcal{T} \mid X)$ by maximising the expected complete-data log-posterior $\mathrm{p}(M, \theta, \beta, \mathcal{T} \mid X, Z)$ iteratively. For convenience we denote by $\mathcal{T}_{\mathsf{b}_i}$ the idiosyncratic precision matrix in batch l, i.e. if $b_{il} = 1$ by τ_{jl}, then the errors are distributed as $\mathsf{e}_i \sim N(0, \mathcal{T}_{\mathsf{b}_i}^{-1})$. We also denote with $\widehat{\Delta} = (\widehat{M}, \widehat{\theta}, \widehat{\beta}, \widehat{\mathcal{T}})$ the current value of the parameters

The E-step takes the expectation of $\log \mathrm{p}(M, \theta, \beta, \mathcal{T} \mid X, Z)$ with respect to $\mathrm{p}(Z \mid \widehat{\Delta}, X)$ Specifically, let

$$
\begin{aligned}
Q(\Delta) =& \mathbb{E}_{z|\cdot}\left[\log \mathrm{p}(M, \theta, \beta, \mathcal{T} \mid X, Z)\right] \\
=& C - \frac{1}{2}\sum_{i=1}^{n}\Big[(\mathsf{x}_i - \theta\mathsf{v}_i - \beta\mathsf{b}_i)^\top \mathcal{T}_{\mathsf{b}_i}(\mathsf{x}_i - \theta\mathsf{v}_i - \beta\mathsf{b}_i) \\
& -2(\mathsf{x}_i - \theta\mathsf{v}_i - \beta\mathsf{b}_i)^\top \mathcal{T}_{\mathsf{b}_i} M \mathbb{E}[\mathsf{z}_i \mid \widehat{\Delta}, X] + \operatorname{tr}\left(M^\top \mathcal{T}_{\mathsf{b}_i} M \mathbb{E}[\mathsf{z}_i\mathsf{z}_i^\top \mid \widehat{\Delta}, X]\right)\Big] \\
& + \sum_{l=1}^{p_b} \frac{n_l + \eta - 2}{2} \log \mid \mathcal{T}_l \mid - \sum_{l=1}^{p_b} \frac{\eta\xi}{2} \operatorname{tr}(\mathcal{T}_l) - \frac{1}{2}\sum_{j=1}^{p} (\theta_j^\top, \beta_j^\top) \frac{1}{\psi} \mathbf{I}(\theta_j, \beta_j),
\end{aligned}
\tag{3.7}
$$

where C is a constant. Expression (3.7) only depends on Z through the conditional posterior mean

$$
\mathbb{E}[\mathsf{z}_i|\widehat{\Delta}, X] = (\mathbf{I}_q + \widehat{M}^\top \widehat{\mathcal{T}_{\mathsf{b}_i}} \widehat{M})^{-1} \widehat{M}^\top \widehat{\mathcal{T}_{\mathsf{b}_i}} (\mathsf{x}_i - \widehat{\theta}\mathsf{v}_i - \widehat{\beta}\mathsf{b}_i) \tag{3.8}
$$

and the conditional second moments

$$
\mathbb{E}[\mathsf{z}_i\mathsf{z}_i^\top \mid \widehat{\Delta}, X] = (\mathbf{I}_q + \widehat{M}^\top \widehat{\mathcal{T}_{\mathsf{b}_i}} \widehat{M})^{-1} + \mathbb{E}[\mathsf{z}_i \mid \widehat{\Delta}, X] \mathbb{E}[\mathsf{z}_i \mid \widehat{\Delta}, X]^\top, \tag{3.9}
$$

where $(\mathbf{I}_q + \widehat{M}^\top \widehat{\mathcal{T}_{\mathsf{b}_i}} \widehat{M})^{-1} = \operatorname{Cov}[\mathsf{z}_i|\widehat{\Delta}, X]$ is the conditional covariance matrix of the latent factors. We emphasise that (3.8) and (3.9) depend on batch-specific precisions $\mathcal{T}_{\mathsf{b}_i}$.

The M-step maximises Q with respect to $M, \theta, \beta, \mathcal{T}$. Setting its partial derivatives to 0 gives the updates

$$
\frac{\partial Q}{\partial M} = -\frac{1}{2}\sum_{i=1}^{n}\left[-2\widehat{\mathcal{T}_{\mathsf{b}_i}}(\mathsf{x}_i - \widehat{\theta}\mathsf{v}_i - \widehat{\beta}\mathsf{b}_i)\mathbb{E}[\mathsf{z}_i^\top \mid \widehat{\Delta}, X] + 2\widehat{\mathcal{T}_{\mathsf{b}_i}}\widehat{M}\mathbb{E}[\mathsf{z}_i\mathsf{z}_i^\top \mid \widehat{\Delta}, X]\right] = 0 \tag{3.10}
$$

The maximum of the j^{th} row of matrix M can be found solving (3.10) as:

$$
\widehat{m}_j = \left[\sum_{i=1}^{n}\left[\widehat{\tau}_j^\top \mathsf{b}_i (x_{ij} - \widehat{\theta}v_{ij} - \widehat{\beta}b_{ij})\mathbb{E}[\mathsf{z}_i^\top \mid \widehat{\Delta}, X]\right]\right]\left[\sum_{i=1}^{n}\left[\widehat{\tau}_j^\top \mathsf{b}_i \mathbb{E}[\mathsf{z}_i\mathsf{z}_i^\top \mid \widehat{\Delta}, X]\right]\right]^{-1} \tag{3.11}
$$

for $j = 1, \ldots, p$.

Maximisation of $\mathcal{T}$ for a fixed batch l is obtained by taking the derivative with respect to $\mathcal{T}_l$

$$
\begin{aligned}
\frac{\partial Q}{\partial \mathcal{T}_l} =& -\frac{1}{2}\sum_{i:\, b_{il}=1}\Big[(\mathsf{x}_i - \widehat{\theta}\mathsf{v}_i - \widehat{\beta}\mathsf{b}_i)(\mathsf{x}_i - \widehat{\theta}\mathsf{v}_i - \widehat{\beta}\mathsf{b}_i)^\top \\
& -2(\mathsf{x}_i - \widehat{\theta}\mathsf{v}_i - \widehat{\beta}\mathsf{b}_i)\mathbb{E}[\mathsf{z}_i \mid \widehat{\Delta}, X]^\top \widehat{M}^\top + \widehat{M}\mathbb{E}[\mathsf{z}_i\mathsf{z}_i^\top \mid \widehat{\Delta}, X]\widehat{M}^\top\Big] \\
& + \frac{n_l + \eta - 2}{2}\widehat{\mathcal{T}_l}^{-1} - \frac{\eta\xi}{2}\mathbf{I}_p = 0.
\end{aligned}
\tag{3.12}
$$

Solving Equation (3.12) and using the diagonal constraint we obtain:

$$\widehat{\mathcal{T}}_l^{-1} = \frac{1}{n_l + \eta - 2}\text{diag}\left\{\sum_{i:\, b_{il}=1}\left(\widetilde{\mathsf{x}}_i\widetilde{\mathsf{x}}_i^\top - 2\widetilde{\mathsf{x}}_i\mathbb{E}[\mathsf{z}_i \mid \widehat{\Delta}, X]^\top\widehat{M}^\top + \widehat{M}\mathbb{E}[\mathsf{z}_i\mathsf{z}_i^\top \mid \widehat{\Delta}, X]\widehat{M}^\top\right) + \eta\xi\mathbf{I}_p\right\} \tag{3.13}$$

with $\widetilde{\mathsf{x}}_i = \mathsf{x}_i - \widehat{\theta}\mathsf{v}_i - \widehat{\beta}\mathsf{b}_i$.

To maximise with respect to (θ, β) we set

$$\frac{\partial Q}{\partial(\theta,\beta)} = -\sum_{i=1}^{n}\left[\widehat{\mathcal{T}}_{\mathsf{b}_i}(\widehat{\theta},\widehat{\beta})(\mathsf{v}_i,\mathsf{b}_i)(\mathsf{v}_i,\mathsf{b}_i)^\top - \widehat{\mathcal{T}}_{\mathsf{b}_i}(\mathsf{x}_i - \widehat{M}\mathbb{E}[\mathsf{z}_i \mid \widehat{\Delta}, X])(\mathsf{v}_i,\mathsf{b}_i)^\top\right] - \frac{1}{\psi}(\widehat{\theta},\widehat{\beta}) = 0 \tag{3.14}$$

Taking the j^{th} row of matrix $(\widehat{\theta},\widehat{\beta})$ and solving Equation (3.14):

$$(\widehat{\theta}_j^\top, \widehat{\beta}_j^\top) = \sum_{i=1}^{n}\left[\widehat{\tau}_j^\top\mathsf{b}_i(x_{ij} - \widehat{m}_j^\top\mathbb{E}[\mathsf{z}_i \mid \widehat{\Delta}, X])(\mathsf{v}_i,\mathsf{b}_i)^\top\right]\left[\sum_{i=1}^{n}\left[\widehat{\tau}_j^\top\mathsf{b}_i(\mathsf{v}_i,\mathsf{b}_i)(\mathsf{v}_i,\mathsf{b}_i)^\top\right] + \frac{1}{\psi}\mathbf{I}\right]^{-1} \tag{3.15}$$

Equation (3.15) has the form of a ridge regression estimator with penalty ψ.

Algorithm 4 summarises the EM algorithm. The stopping criteria is reaching a tolerance ε^* in the log-posterior change, a maximum number of iterations T or a change ε_M^* on the loadings. By default we set $\varepsilon^* = 0.001$, $T = 100$ and $\varepsilon_M^* = 0.05$. Parameter initialisation is an important aspect that helps obtain better local modes and reduce computational time; its discussion is deferred to Section 3.4.3.

Algorithm 4: EM algorithm for factor regression model with uniform p(M)

initialise $\widehat{M} = M^{(0)}$, $\widehat{\theta} = \theta^{(0)}$, $\widehat{\beta} = \beta^{(0)}$, $\widehat{\mathcal{T}_{\mathsf{b}_i}} = \mathcal{T}_{\mathsf{b}_i}^{(0)}$

while $\varepsilon > \varepsilon^*$, $\varepsilon_M > \varepsilon_M^*$ *and* $t < T$ **do**

E-step:

Latent factors: $\mathbb{E}[\mathsf{z}_i|\widehat{\Delta}, X] = (\mathbf{I}_q + \widehat{M}^\top\widehat{\mathcal{T}_{\mathsf{b}_i}}\widehat{M})^{-1}\widehat{M}^\top\widehat{\mathcal{T}_{\mathsf{b}_i}}(\mathsf{x}_i - \widehat{\theta}\mathsf{v}_i - \widehat{\beta}\mathsf{b}_i)$

M-step:

Loadings: $\widehat{m}_j = \left[\sum_{i=1}^n\left(\widehat{\tau}_j^\top\mathsf{b}_i\widetilde{x}_{ij}\mathbb{E}[\mathsf{z}_i^\top \mid \widehat{\Delta}, X]\right)\right]\left[\sum_{i=1}^n\left(\widehat{\tau}_j^\top\mathsf{b}_i\mathbb{E}[\mathsf{z}_i\mathsf{z}_i^\top \mid \widehat{\Delta}, X]\right)\right]^{-1}$

Variances: $\widehat{\mathcal{T}}_l^{-1} = \frac{1}{n_l+\eta-2}\text{diag}\left\{\sum_{i:\, b_{il}=1}\left(\widetilde{\mathsf{x}}_i\widetilde{\mathsf{x}}_i^\top - 2\widetilde{\mathsf{x}}_i\mathbb{E}[\mathsf{z}_i \mid \widehat{\Delta}, X]^\top\widehat{M}^\top + \widehat{M}\mathbb{E}[\mathsf{z}_i\mathsf{z}_i^\top \mid \widehat{\Delta}, X]\widehat{M}^\top\right) + \eta\xi\mathbf{I}_p\right\}$

Coefficients: $(\widehat{\theta}_j^\top, \widehat{\beta}_j^\top) = \sum_{i=1}^n\left[\widehat{\tau}_j^\top\mathsf{b}_i(x_{ij} - \widehat{m}_j^\top\mathbb{E}[\mathsf{z}_i \mid \widehat{\Delta}, X])(\mathsf{v}_i,\mathsf{b}_i)^\top\right]\left[\sum_{i=1}^n\left[\widehat{\tau}_j^\top\mathsf{b}_i(\mathsf{v}_i,\mathsf{b}_i)(\mathsf{v}_i,\mathsf{b}_i)^\top\right] + \frac{1}{\psi}\mathbf{I}\right]^{-1}$

set $\Delta^{(t+1)} = \widehat{\Delta}$ and $M^{(t+1)} = \widehat{M}$

compute $\varepsilon = Q(\Delta^{t+1}) - Q(\Delta^t)$, $\varepsilon_M = \max||m_{jk}^{(t+1)} - m_{jk}^{(t)}||$ and $t = t + 1$

end

3.4.2 EM algorithm for spike-and-slab priors

The algorithm is derived analogously to Section 3.4.1. The expected complete-data log-posterior can be split into $Q(\Delta) = C + Q_1(\theta, M, \beta, \mathcal{T}) + Q_2(\zeta)$, where

$$\begin{aligned} Q_1(\theta, M, \beta, \mathcal{T}) = &-\frac{1}{2}\sum_{i=1}^{n}\Big[(\mathsf{x}_i - \theta\mathsf{v}_i - \beta\mathsf{b}_i)^\top \mathcal{T}_{\mathsf{b}_i}(\mathsf{x}_i - \theta\mathsf{v}_i - \beta\mathsf{b}_i) \\ &-2(\mathsf{x}_i - \theta\mathsf{v}_i - \beta\mathsf{b}_i)^\top \mathcal{T}_{\mathsf{b}_i} M\mathbb{E}[\mathsf{z}_i \mid \widehat{\Delta}, X] + \operatorname{tr}\left(M^\top \mathcal{T}_{\mathsf{b}_i} M\mathbb{E}[\mathsf{z}_i\mathsf{z}_i^\top \mid \widehat{\Delta}, X]\right)\Big] \\ &+\sum_{l=1}^{p_b}\frac{n_l + \eta - 2}{2}\log \mid \mathcal{T}_l \mid - \sum_{l=1}^{p_b}\frac{\eta\xi}{2}\operatorname{tr}(\mathcal{T}_l) \\ &-\frac{1}{2}\sum_{j=1}^{p}(\theta_j, \beta_j)^\top \frac{1}{\psi}\mathbf{I}(\theta_j, \beta_j) + \sum_{j=1}^{p}\sum_{k=1}^{q}\mathbb{E}_{\gamma\mid\cdot}\left[\log \mathrm{p}(m_{jk} \mid \gamma_{jk}, \lambda_0, \lambda_1)\right], \end{aligned} \tag{3.16}$$

$$Q_2(\zeta) = \sum_{j=1}^{p}\sum_{k=1}^{q}\log\left(\frac{\zeta_k}{1-\zeta_k}\right)\mathbb{E}[\gamma_{jk} \mid \widehat{\Delta}] + \sum_{k=1}^{q}\left((\frac{a_\zeta}{k} - 1)\log(\zeta_k) + (p + b_\zeta - 1)\log(1-\zeta_k)\right). \tag{3.17}$$

with C a constant and $\mathbb{E}[\mathsf{z}_i \mid \widehat{\Delta}, X]$ and $\mathbb{E}[\mathsf{z}_i\mathsf{z}_i^\top \mid \widehat{\Delta}, X]$ as in (3.8) and (3.9).

$Q_1(\theta, M, \beta, \mathcal{T})$ resembles the E-step for the flat prior in Section 3.4.1, plus an extra conditional expectation $\mathbb{E}_{\gamma\mid\cdot}\left[\log \mathrm{p}(m_{jk} \mid \gamma_{jk}, \lambda_0, \lambda_1)\right]$. $Q_2(\zeta)$ arises from the Beta-Binomial prior on γ_{jk} and the $\mathbb{E}[\gamma_{jk} \mid \widehat{\Delta}] = \mathrm{p}(\gamma_{jk} = 1 \mid \widehat{\Delta}) = \widehat{p}_{jk}$ have been already computed for the following priors: Normal-SS in (2.25), Laplace-SS in (2.33), MOM-SS in (2.38), and Laplace-MOM-SS in (2.43).

In the M-step we maximise Q_1 w.r.t. $(\theta, M, \beta, \mathcal{T})$, this can be done in a completely independent fashion from optimising Q_2 w.r.t. ζ.

The main difference between the local and non-local priors lies in updating the loadings. We discuss these separately for each prior later in this section.

The updates for the precision $\mathcal{T}_l$, the regression parameters (θ, β) and the weights ζ_k are given in Equations (3.13), (3.15) and (2.30) respectively.

Algorithm 5 summarises the EM procedure. It is initialised with the two-stage least-squares method described in Section 3.4.3 and $\zeta_k = 0.5$ for $k = 1, \ldots, q$. The stopping criteria are as in Algorithm 4. The different updates for M are outlined below, separately for each prior specification.

Algorithm 5: EM algorithm for factor regression model with spike-and-slab p(M)

initialise $\widehat{M} = M^{(0)}$, $\widehat{\theta} = \theta^{(0)}$, $\widehat{\beta} = \beta^{(0)}$, $\widehat{\mathcal{T}_{\mathrm{b}_i}} = \mathcal{T}_{\mathrm{b}_i}^{(0)}$, $\widehat{\zeta} = \zeta^{(0)}$

while $\varepsilon > \varepsilon^*$, $\varepsilon_M > \varepsilon_M^*$ *and* $t < T$ **do**

E-step:

Latent factors: $\mathbb{E}[\mathrm{z}_i|\widehat{\Delta}, X] = (\mathrm{I}_q + \widehat{M}^\top \widehat{\mathcal{T}_{\mathrm{b}_i}} \widehat{M})^{-1} \widehat{M}^\top \widehat{\mathcal{T}_{\mathrm{b}_i}} (\mathrm{x}_i - \widehat{\theta}\mathrm{v}_i - \widehat{\beta}\mathrm{b}_i)$

Latent indicators$^+$: $\mathbb{E}[\gamma_{jk} \mid \widehat{\Delta}] = \widehat{p}_{jk}$

M-step:

Loadings$^+$: $\widehat{m}_{jk} = \arg\max_{m_{jk}} Q_1(\widehat{\Delta})$

Variances: $\widehat{\tau}_l^{-1} = \frac{1}{n_l + \eta - 2} \mathrm{diag}\left\{ \sum_{i:\, \mathrm{b}_{il}=1} \left(\widetilde{x}_i \widetilde{x}_i^\top - 2\widetilde{x}_i \mathbb{E}[\mathrm{z}_i \mid \widehat{\Delta}, X]^\top \widehat{M}^\top + \widehat{M} \mathbb{E}[\mathrm{z}_i \mathrm{z}_i^\top \mid \widehat{\Delta}, X] \widehat{M}^\top \right) + \eta \xi \mathrm{I}_p \right\}$

Coefficients: $(\widehat{\theta}_j^\top, \widehat{\beta}_j^\top) = \sum_{i=1}^n \left[\widehat{\tau}_j^\top \mathrm{b}_i (x_{ij} - \widehat{m}_j^\top \mathbb{E}[\mathrm{z}_i \mid \widehat{\Delta}, X])(\mathrm{v}_i, \mathrm{b}_i)^\top \right] \left[\sum_{i=1}^n \left[\widehat{\tau}_j^\top \mathrm{b}_i (\mathrm{v}_i, \mathrm{b}_i)(\mathrm{v}_i, \mathrm{b}_i)^\top \right] + \frac{1}{\psi} \mathrm{I} \right]^{-1}$

Weights: $\widehat{\zeta}_k = \frac{\sum_{j=1}^p \widehat{p}_{jk} + \frac{a_\zeta}{k} - 1}{\frac{a_\zeta}{k} + b_\zeta + p - 1}$

set $\Delta^{(t+1)} = \widehat{\Delta}$ and $M^{(t+1)} = \widehat{M}$

compute $\varepsilon = Q(\Delta^{t+1}) - Q(\Delta^t)$, $\varepsilon_M = \max ||m_{jk}^{(t+1)} - m_{jk}^{(t)}||$ and $t = t + 1$

end

$^+$ see Section 3.4.2 for details.

Normal-SS prior

Let $d_{jk} = [(1 - \gamma_{jk})\lambda_0 + \gamma_{jk}\lambda_1]^{-1}$. In Expression (3.16), under a Normal-SS prior

$$\mathbb{E}_{\gamma|\cdot}\left[\log \mathrm{p}(m_{jk} \mid \gamma_{jk}, \lambda_0, \lambda_1)\right] \propto -\frac{1}{2} m_{jk}^2 \mathbb{E}\left[d_{jk} \mid \widehat{\Delta}\right] = -\frac{1}{2} m_{jk}^2 \left[\frac{1 - \widehat{p}_{jk}}{\lambda_0} + \frac{\widehat{p}_{jk}}{\lambda_1}\right] \tag{3.18}$$

where $\widehat{p}_{jk}$ is as in (2.25).

Setting to 0 the partial derivative with respect to M gives:

$$\begin{aligned} \frac{\partial Q}{\partial M} = &- \frac{1}{2} \sum_{i=1}^n \left[-2\widehat{\mathcal{T}}_{\mathrm{b}_i} (\mathrm{x}_i - \widehat{\theta}\mathrm{v}_i - \widehat{\beta}\mathrm{b}_i) \mathbb{E}[\mathrm{z}_i^\top \mid \widehat{\Delta}, X] + 2\widehat{\mathcal{T}}_{\mathrm{b}_i} \widehat{M} \mathbb{E}[\mathrm{z}_i \mathrm{z}_i^\top \mid \widehat{\Delta}, X] \right] \\ &- \widehat{M} \circ \mathbb{E}[D_\gamma \mid \widehat{\Delta}] = 0, \end{aligned} \tag{3.19}$$

with $D_\gamma \in \mathbb{R}^{p \times q}$ with the element jk being d_{jk} and $A \circ B$ being the Hadamard (element-wise) product of two matrices A and B. Taking the j^{th} row of matrix M and solving equation (3.19) we obtain:

$$\widehat{m}_j = \left[\sum_{i=1}^n \left(\widehat{\tau}_j^\top \mathrm{b}_i \widetilde{x}_{ij} \mathbb{E}[\mathrm{z}_i^\top \mid \widehat{\Delta}, X] \right) \right] \left[\mathrm{diag}\{\mathbb{E}[d_{j1} \mid \widehat{\Delta}], \ldots, \mathbb{E}[d_{jq} \mid \widehat{\Delta}]\} + \sum_{i=1}^n \left(\widehat{\tau}_j^\top \mathrm{b}_i \mathbb{E}[\mathrm{z}_i \mathrm{z}_i^\top \mid \widehat{\Delta}, X] \right) \right]^{-1}, \tag{3.20}$$

for $j = 1, \ldots, p$, where $\widetilde{x}_{ij} = x_{ij} - \theta v_{ij} - \beta b_{ij}$.

MOM-SS prior

For MOM-SS

$$\mathbb{E}_{\gamma|\cdot}\left[\log \mathrm{p}(m_{jk} \mid \gamma_{jk}, \widetilde{\lambda}_0, \widetilde{\lambda}_1)\right] \propto -\frac{1}{2} m_{jk}^2 \left[\frac{1-\widehat{p}_{jk}}{\widetilde{\lambda}_0} + \frac{\widehat{p}_{jk}}{\widetilde{\lambda}_1}\right] + \widehat{p}_{jk} \log(m_{jk}^2). \tag{3.21}$$

where $\widehat{p}_{jk}$ is given in (2.38)

For the M-step, we use a coordinate descent algorithm (CDA), doing successive univariate optimisation on (3.16) with respect to each m_{jk}.

Viewed as a function of only m_{jk}, it is possible to express $Q_1(m_{jk})$ as

$$Q_1(m_{jk}) = \frac{a}{2} m_{jk}^2 + b m_{jk} + \frac{c}{2} \log(m_{jk}^2), \tag{3.22}$$

where

$$\begin{aligned} a &= -\left(\left[\frac{1-\widehat{p}_{jk}}{\widetilde{\lambda}_0} + \frac{\widehat{p}_{jk}}{\widetilde{\lambda}_1}\right] + \sum_{i=1}^{n} \widehat{\tau}_j^\top \mathrm{b}_i \mathbb{E}[z_{ik} z_{ik}^\top \mid \widehat{\Delta}, X]\right) \\ b &= \sum_{i=1}^{n} \left[\widehat{\tau}_j^\top \mathrm{b}_i (x_{ij} - \widehat{\theta} v_{ij} - \widehat{\beta} b_{ij}) \mathbb{E}[z_{ik} \mid \widehat{\Delta}, X] - \sum_{r \neq k}^{q} \widehat{m}_{jr} \widehat{\tau}_j^\top \mathrm{b}_i \mathbb{E}[z_{ir} z_{ik}^\top \mid \widehat{\Delta}, X]\right] \\ c &= 2\widehat{p}_{jk} \end{aligned} \tag{3.23}$$

Define

$$\underline{\mathrm{m}}_{jk} = \frac{-b - \sqrt{b^2 - 4ac}}{2a}$$

and

$$\overline{m}_{jk} = \frac{-b + \sqrt{b^2 - 4ac}}{2a}.$$

It follows from Lemma 2.3 that the global maximum of (3.22) is

$$\arg\max_{m_{jk}} f(m_{jk}) = \begin{cases} \underline{\mathrm{m}}_{jk} & \text{if } b > 0 \\ \overline{m}_{jk} & \text{if } b < 0 \\ \underline{\mathrm{m}}_{jk} = \overline{m}_{jk} & \text{if } b = 0 \end{cases}$$

as $a < 0$ and $c < 0$.

Laplace-SS and Laplace-MOM-SS priors

Akin to the MOM-SS, we can express

$$Q_1(m_{jk}) = \frac{a}{2}m_{jk}^2 + bm_{jk} + c|m_{jk}| + \frac{d}{2}\log(m_{jk}^2)$$

$$\begin{aligned}
a &= -\sum_{i=1}^{n} \widehat{\tau}_j^\top \mathrm{b}_i \mathbb{E}[z_{ik} z_{ik}^\top \mid \widehat{\Delta}, X] \\
b &= \sum_{i=1}^{n} \left[\widehat{\tau}_j^\top \mathrm{b}_i (x_{ij} - \widehat{\theta} v_{ij} - \widehat{\beta} b_{ij}) \mathbb{E}[z_{ik} \mid \widehat{\Delta}, X] - \sum_{r \neq k}^{q} \widehat{m}_{jr} \widehat{\tau}_j^\top \mathrm{b}_i \mathbb{E}[z_{ir} z_{ik}^\top \mid \widehat{\Delta}, X] \right] \\
c &= -\left[\frac{1 - \widehat{p}_{jk}}{\lambda_0} + \frac{\widehat{p}_{jk}}{\lambda_1} \right] \\
d &= \begin{cases} 0 & \text{for Laplace-SS} \\ 2\widehat{p}_{jk} & \text{for Lapace-MOM-SS} \end{cases}
\end{aligned} \tag{3.24}$$

for $j = 1, \ldots, p$ and where $\widehat{p}_{jk}$ is as in (2.33) and (2.43) for Laplace-SS and Laplace-MOM-SS respectively.

For Laplace-SS, define

$$m_{jk}^+ = \frac{-(b + c)}{a}$$

and

$$m_{jk}^- = \frac{-(b - c)}{a},$$

with $a < 0$ and $c < 0$. The global maximum is then

$$\arg\max_{m_{jk}} f(m_{jk}) = \begin{cases} m_{jk}^+ & \text{if } b > -c \\ m_{jk}^- & \text{if } b < c \\ m_{jk}^+ = m_{jk}^- & \text{if } c < b < -c \end{cases}$$

with $a < 0$ and $c < 0$, as in Lemma 2.1.

Finally for the Laplace-MOM-SS, we note that when $m_{jk} = 0$, $Q_1(m_{jk} = 0) = -\infty$. Thus the solution for m_{jk} is given by setting $\frac{\partial Q_1}{\partial m_{jk}} = 0$. Define

$$m_{jk}^+ = \frac{-(b + c) - \sqrt{(b + c)^2 - 4ad}}{2a}$$

and

$$m_{jk}^{-} = \frac{-(b-c) + \sqrt{(b-c)^2 - 4ad}}{2a}.$$

The global maximum follows from Lemma 2.4 as

$$\arg\max_{m_{jk}} f(m_{jk}) = \begin{cases} m_{jk}^{+} & \text{if } b > 0 \\ m_{jk}^{-} & \text{if } b < 0 \\ m_{jk}^{+} = m_{jk}^{-} & \text{if } b = 0 \end{cases}$$

where $a < 0$, $c < 0$ and $d > 0$.

We highlight that if either x_i or v_i is continuous, the event of $b = 0$ has zero probability. If both x_i and v_i are discrete and in presence of the rare event of $b = 0$, then the sign of the update for m_{jk} is set to the signs of its current value.

3.4.3 Initialisation of parameters

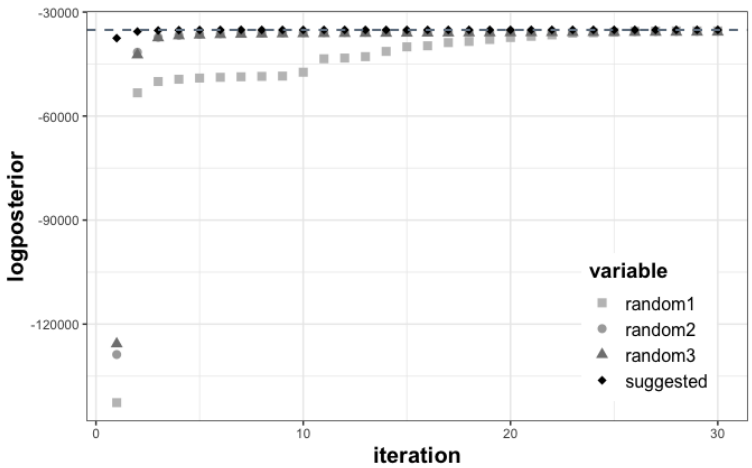


Figure 22. Comparison of the log-posterior convergence at four different initialisations of parameters. The suggested initialisation in squared shape and the true log-posterior in dotted line.

As mentioned in Section 2.11, the EM algorithm can be sensitive to parameter initialisation. To fix this, we modify the two initialisations proposed in Section 2.11 by adding a simple regression step to each of them:

(i) A simple two-step least-squares.

- Step 1: initialise $(\theta^{(0)}, \beta^{(0)}) = [(V, B)^\top (V, B)]^{-1} (V, B)^\top X$.

- Step 2: let $\widehat{E} = X - (V\theta^{(0)\top} + B\beta^{(0)\top})$; consider the eigendecomposition of $\frac{1}{n}\widehat{E}^\top\widehat{E}$ where $l_1 \geq l_2 \geq l_q$ are the eigenvalues and $u_1, \ldots, u_q$ the eigenvectors; set $M^{(0)} = [\sqrt{l_1}u_1 \mid \cdots \mid \sqrt{l_q}u_q]$ and $\mathcal{T}_l^{(0)} = [\text{diag}\{\frac{1}{n}\widehat{E}^\top\widehat{E} - M^{(0)}M^{(0)\top}\}]^{-1}$ for $l = 1, \ldots, p_b$.

(ii) The rotated least-squares adds an extra step.

- Step 3: varimax rotation for the loadings obtained in Step 2.

3.4.4 Post-processing for model selection, dimensionality reduction and normalised data visualisation

As explained in Section 2.12, under Normal-SS, MOM-SS and Laplace-MOM-SS the resulting $\widehat{M}$ are not sparse, thus we applied the two straightforward post-processing steps proposed in Section 2.12.

The two possible post-processing options combined with the two initialisation alternatives gives us four possible solutions for $\widehat{M}$. In order to select the best alternative for our models, we extended the weighted 10-fold cross-validation in Algorithm 3 to Algorithm 6: here, batches with higher variance receive lower weight and the selected model is the one with smallest weighted cross validation reconstruction error.

Further, we perform the left-ordering function in Definition 2.5 on the inclusion matrix to ease the interpretation of latent factors.

Latent factors are also post-processed for data visualisation purposes. The aim of this is to obtain new standardised factors $\widetilde{\mathsf{z}}_i = [\text{Cov}(\mathsf{z}_i \mid \widehat{\Delta}, X)]^{-1}\mathbb{E}[\mathsf{z}_i \mid \widehat{\Delta}, X]$, with $\text{Cov}(\mathsf{z}_i \mid \widehat{\Delta}, X) = (\mathbf{I}_q + \widehat{M}^\top\widehat{\mathcal{T}_{\mathsf{b}_i}}\widehat{M})^{-1}$, whose covariance does not depend on their batch.

3.5 Results

To quantify the effectiveness of our approach we first simulated datasets. Akin to Section 2.13, we compare our methods with the FastBFA of Ročková and George (2017) and the LASSO-BIC of Hirose and Yamamoto (2015) in addition to the ComBat empirical Bayes batch effect correction of Johnson et al. (2007) for scenarios with batch effects, doing an MLE estimation of the factor analysis model (ComBat-MLE). The **R** code for our model with batch effect correction is available at `https://github.com/AleAviP/BFR.BE`. We used the package `sva 3.26.0` for ComBat (Leek et al., 2017). Hyper-parameters for the Normal-SS and MOM-SS were set as in Section 2.10 and for FastBFA and LASSO-BIC as in Section 2.13. Finally, for scenarios with batch effects, we adjusted the data via a ComBat correction and performed a Factor Analysis via EM algorithm to maximise likelihood with the `fa.em` function in the `cate` package (Wang and Zhao, 2015).

Algorithm 6: Weighted 10-fold cross-validation for Bayesian factor regression

initialise $\varepsilon_X = 0$
set 10 random cross-validation subsets of
Observations: $\{\mathsf{x}^{[1]}, \ldots, \mathsf{x}^{[10]}\} \in \mathbb{R}^{\frac{n}{10} \times p}$
Covariates: $\{\mathsf{v}^{[1]} \ldots, \mathsf{v}^{[10]}\} \in \mathbb{R}^{\frac{n}{10} \times p_v}$
Batches: $\{\mathsf{b}^{[1]}, \ldots, \mathsf{b}^{[10]}\} \in \mathbb{R}^{\frac{n}{10} \times p_b}$
for $r \leftarrow 1, \ldots, 10$ **do**
  set: Cross-validation subsets
  $\widetilde{\mathsf{x}} := (\mathsf{x}^{[1]}, \ldots, \mathsf{x}^{[r-1]}, \mathsf{x}^{[r+1]}, \ldots, \mathsf{x}^{[10]}), \widetilde{\mathsf{v}} := (\mathsf{v}^{[1]}, \ldots, \mathsf{v}^{[r-1]}, \mathsf{v}^{[r+1]}, \ldots, \mathsf{v}^{[10]})$
  and $\widetilde{\mathsf{b}} := (\mathsf{b}^{[1]}, \ldots, \mathsf{b}^{[r-1]}, \mathsf{b}^{[r+1]}, \ldots, \mathsf{b}^{[10]})$,
  compute: EM algorithm
    input: $\widetilde{\mathsf{x}}, \widetilde{\mathsf{v}}, \widetilde{\mathsf{b}}$
    output: $\widehat{M}, \widehat{\mathcal{T}_{\mathsf{b}_i}}, \widehat{\theta}, \widehat{\beta}, \widehat{\zeta}$
  set: Test factors $\widehat{\mathsf{z}}_i = (\mathbf{I}_q + \widehat{M}^\top \widehat{\mathcal{T}_{\mathsf{b}_i}} \widehat{M})^{-1} \widehat{M}^\top \widehat{\mathcal{T}_{\mathsf{b}_i}} (\mathsf{x}^{[r]} - \widehat{\theta}\mathsf{v}^{[r]} - \widehat{\beta}\mathsf{b}^{[r]})$
  compute $\varepsilon_X = \varepsilon_X + \sum_i || \left[\mathsf{x}_i^{[r]} - (\widehat{\theta}\mathsf{v}_i^{[r]} + \widehat{M}\widehat{\mathsf{z}}_i + \widehat{\beta}\mathsf{b}_i^{[r]}) \right] \widehat{\mathcal{T}_{\mathsf{b}_i}} ||_F$
end
set $\varepsilon_X = \frac{\varepsilon_X}{10}$

We simulated data from two different data-generating truths:

(i) dense loadings matrix with a grid of elements set uniformly between $(-1, 1)$

(ii) truly sparse loadings with a banded-diagonal structure with $m_{jk} = 1$ for the non-zero elements.

In both, the truth was set to $q^* = 10$ factors (see Figure 14).

We evaluate our method in our main setting of interest where there are mean and variance batch effects.

We simulated data with a mean and variance batch effect, $\mathsf{x}_i = \theta^* \mathsf{v}_i + M^* \mathsf{z}_i + \beta^* \mathsf{b}_i + \mathsf{e}_i$, sample size $n = 200$ and growing $p = 250$ or $p = 500$. We set $q^* = 10$, $p_v = 1$ and $p_b = 2$ batches and considered the truly sparse and dense loadings M^* in Figure 14. Factors z_i were drawn from $N(0, \mathbf{I}_q)$, errors e_i from $N(0, \mathcal{T}_{\mathsf{b}_i}^{-1})$, where $\tau_{j1}^{-1} = 0.5$ and $\tau_{j2}^{-1} = 1.5\tau_{j1}^{-1}$ for $j = 1, \ldots, p$; v_i from a continuous Uniform(0,3) and b_i from a discrete Uniform$\{0,1\}$. We set the first $p/2$ values of $\theta^* \in \mathbb{R}^p$ to -2 and the other $p/2$ to 2 and $\beta_{j1}^* = 0$, $\beta_{j2}^* = 2$ for $j = 1, \ldots, p$ we fixed to 2 for the first batch and 0 for the second. We compared our models with FastBFA and LASSO-BIC without batch effect correction for illustration of the importance of a proper mean and variance batch effect adjustment; and with empirical Bayes batch effect correction, ComBat, followed with an MLE estimation of the parameters ComBat-MLE. We emphasise that, to our knowledge, there are no other model-based methods to learn the latent structure while correcting for non-biological variance simultaneously. Thus, this is not a fair comparison with FastBFA and LASSO-BIC but rather

an illustration of how much inference when not properly accounting for batch effects in a model-based manner.

To quantify the performance of the methods, Tables 3 and 4 show the estimated latent cardinality $\widehat{q}$, the number of non-zero loadings $||\widehat{M}||_0 = \sum_{j,k} \mathbb{1}(\widehat{m}_{jk} \neq 0)$, the Frobenius norm (F.N.) between the true expected value and its reconstruction $||E[X] - \widehat{E}[X]||_F = ||(ZM^\top + V\theta^\top + B\beta^\top) - (\mathbb{E}[Z \mid \widehat{\Delta}, X]\widehat{M}^\top + V\widehat{\theta}^\top + B\widehat{\beta}^\top)||_F$ and between the true and reconstructed factors and loadings $||ZM^\top - \mathbb{E}[Z \mid \widehat{\Delta}, X]\widehat{M}^\top||_F$, and the number of iterations until convergence. We display the mean across 100 independent simulations (See Appendix A).

3.5.1 Dense loadings

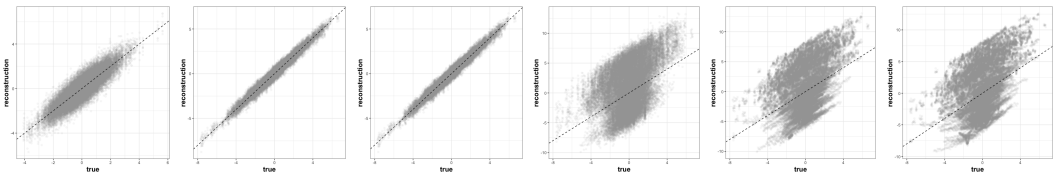
Table 3: Synthetic data with batch effects for $n = 200$, $q^* = 10$, $p = 250$ or 500 parameters, dense loadings M^*.

	$p = 250$					$p = 500$				
Model	$\widehat{q}$	$\|\|\widehat{M}\|\|_0$	$\|\|\mathbb{E}[X] - \widehat{\mathbb{E}}[X]\|\|_F$	$\|\|ZM^\top - \mathbb{E}[Z \mid \cdot]\widehat{M}^\top\|\|_F$	it	$\widehat{q}$	$\|\|\widehat{M}\|\|_0$	$\|\|\mathbb{E}[X] - \widehat{\mathbb{E}}[X]\|\|_F$	$\|\|ZM^\top - \mathbb{E}[Z \mid \cdot]\widehat{M}^\top\|\|_F$	it
	$q = 10$									
Flat	10.0	2500.0	56.5	88.2	4.4	10.0	5000.0	71.9	120.2	4.0
Normal-SS	10.0	727.6	**54.0**	**83.9**	8.0	10.0	1398.7	**68.6**	**116.5**	4.5
MOM-SS	10.0	1097.3	55.1	84.6	15.2	10.0	1257.5	70.1	127.4	81.1
ComBat-MLE	10.0	2500.0	178.5	810.2	3.1	10.0	5000.0	249.2	1144.9	3.2
FastBFA	10.0	1153.0	89.0	834.5	12.3	10.0	2343.1	106.6	1182.6	10.9
LASSO-BIC	10.0	2109.9	99.2	833.1	NA	10.0	4377.1	118.1	1182.9	NA
	$q = 100$									
Flat	100.0	25000.0	140.7	157.6	5.0	100.0	50000.0	208.8	231.2	10.7
Normal-SS	29.7	983.5	87.4	111.2	6.3	10.0	2725.1	**73.4**	**119.8**	5.6
MOM-SS	10.0	1216.7	**57.4**	**87.7**	7.2	10.0	2293.5	74.0	120.4	6.3
ComBat-MLE	100.0	25000.0	70.6	822.6	33.8	100.0	50000.0	123.3	1161.0	14.8
FastBFA	35.3	1285.5	79.3	826.7	19.6	59.9	2589.0	126.8	1181.6	12.5
LASSO-BIC	12.9	1579.6	59.4	827.8	NA	11.1	2939.6	75.8	1171.2	NA

Firstly, we considered the scenario when one correctly guesses $q = 10$ and loadings are truly dense so that sparse priors could potentially lead to poor estimations. Table 3 shows the results. MOM-SS and Normal-SS achieved similar performance as the case without batch effect and similar results were observed for the $q = 100$ case. MOM-SS estimated correctly the latent cardinality $q^* = 10$ and achieved a small estimation error for $\mathbb{E}[X]$. Figure 23 compares the true $ZM^{*\top}$ against their reconstruction $\mathbb{E}[Z \mid \widehat{\Delta}, X]\widehat{M}^\top$. Figures 24 and 25 display a graphical representation of $\widehat{M}$, $\widehat{\mathrm{Cov}}(\mathrm{x}_i)^{-1}$ and $\widehat{\gamma}$ setting $q = 100$.

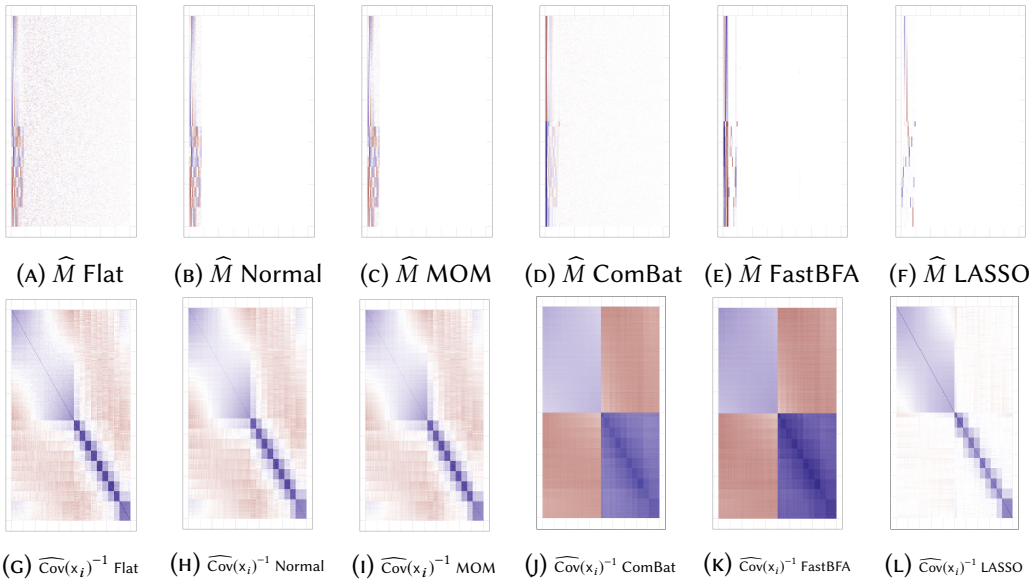
3.5.2 Truly sparse loadings

Secondly, we studied the scenario with truly sparse factor loadings M^*. Table 4 provides a summary of the results, showing that MOM-SS achieved a small estimation error for the mean and was effective in estimating $q^* = 10$. LASSO-BIC had a small estimation error of the mean, although solutions were generally less sparse in the number of non-zero loadings.



Flat Normal-SS MOM-SS ComBat-MLE FastBFA LASSO

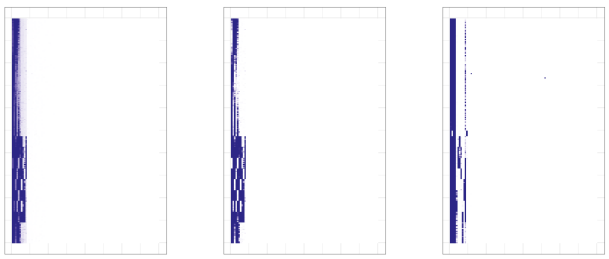
Figure 23. Scatterplots comparing $ZM^\top$ vs. $\mathbb{E}[Z \mid \widehat{\Delta}, X]\widehat{M}^\top$ between the different models under dense loadings M with $q = 100$ in simulations with batch effect.



(A) $\widehat{M}$ Flat (B) $\widehat{M}$ Normal (C) $\widehat{M}$ MOM (D) $\widehat{M}$ ComBat (E) $\widehat{M}$ FastBFA (F) $\widehat{M}$ LASSO

(G) $\widehat{\mathrm{Cov}(x_i)}^{-1}$ Flat (H) $\widehat{\mathrm{Cov}(x_i)}^{-1}$ Normal (I) $\widehat{\mathrm{Cov}(x_i)}^{-1}$ MOM (J) $\widehat{\mathrm{Cov}(x_i)}^{-1}$ ComBat (K) $\widehat{\mathrm{Cov}(x_i)}^{-1}$ FastBFA (L) $\widehat{\mathrm{Cov}(x_i)}^{-1}$ LASSO

Figure 24. Heatmaps of loadings and covariance (red denotes large negative values, blue large positive values, white denotes zero).



(A) $\widehat{\gamma}$ Normal (B) $\widehat{\gamma}$ MOM (C) $\widehat{\gamma}$ FastBFA

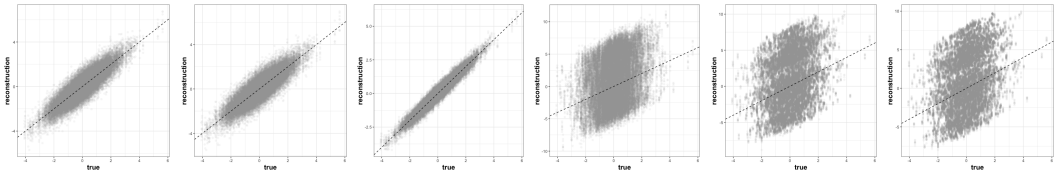
Figure 25. Heatmaps of inclusion probability (white denotes 0, dark blue denotes 1).

It is important to highlight that even though ComBat-MLE, FastBFA and LASSO-BIC achieved a precise reconstruction of $\mathbb{E}[X] = ZM^\top + V\theta^\top + B\beta^\top$, for purposes of

dimensionality reduction the estimates of $ZM^\top$ are less precise, as shown in Tables 3 and 4 and Figures 23 and 26 (right panels). This is due to the following fact: by omitting the covariates V and the batches B from the model, the effect of (V, B) on $\mathbb{E}[X]$ is captured by Z, leading to reasonably accurate reconstructions of $\mathbb{E}[X]$ but poor reconstructions of $ZM^\top$. Furthermore, the estimated covariance of the model displayed in the heatmap in Figures 24 (j)-(l) and 27 (j)-(l) are nowhere close to the generating truth. We remark that for FastBFA and LASSO-BIC these results mainly highlight that one should take into account batch effects. For Combat-MLE they highlight the limitations of using two-step procedures relative to a joint estimation of the factor model and batch effects

Table 4: Synthetic data with batch effects for $n = 200$, $q^* = 10$, $p = 250$ or 500 parameters, truly sparse loadings M^*.

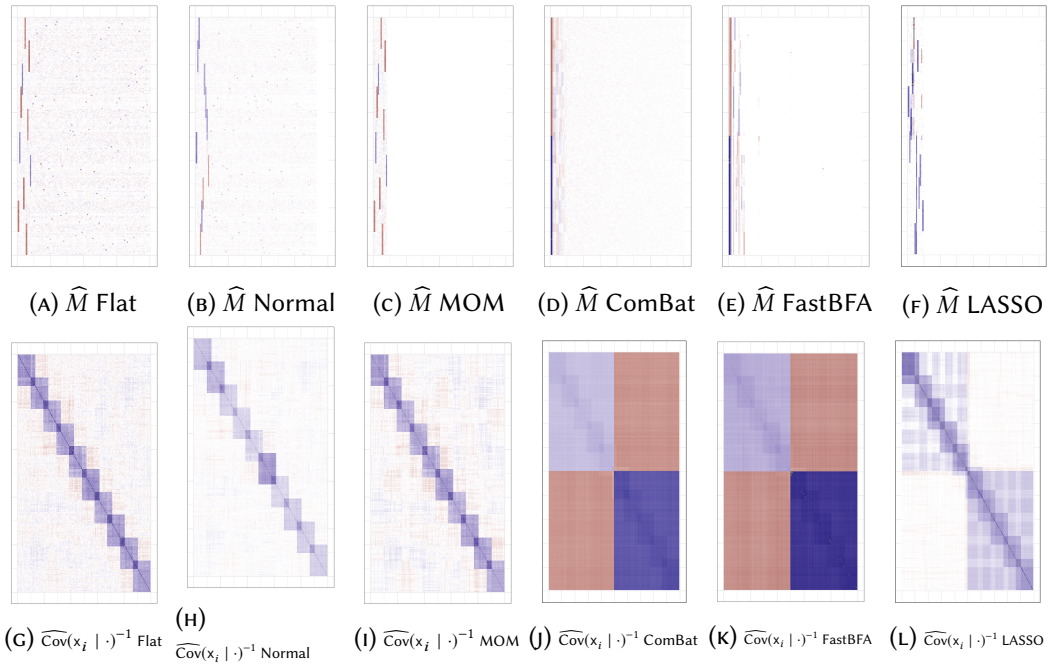
	$p = 250$					$p = 500$				
Model	$\widehat{q}$	$\|\|\widehat{M}\|\|_0$	$\|\|\mathbb{E}[X] - \widehat{\mathbb{E}}[X]\|\|_F$	$\|\|ZM^\top - \mathbb{E}[Z \mid \cdot]\widehat{M}^\top\|\|_F$	it	$\widehat{q}$	$\|\|\widehat{M}\|\|_0$	$\|\|\mathbb{E}[X] - \widehat{\mathbb{E}}[X]\|\|_F$	$\|\|ZM^\top - \mathbb{E}[Z \mid \cdot]\widehat{M}^\top\|\|_F$	it
				$q = 10$						
Flat	10.0	2500.0	49.7	68.5	4.1	10.0	5000.0	60.8	90.7	4.1
Normal-SS	10.0	330.0	45.7	58.7	4.9	10.0	650.0	**55.9**	77.0	4.1
MOM-SS	10.0	330.0	**45.5**	**57.8**	5.4	10.0	650.0	56.0	**76.6**	4.1
ComBat-MLE	10.0	2500.0	171.4	807.8	2.0	10.0	5000.0	244.5	1140.3	1.0
FastBFA	10.0	817.1	78.1	832.1	9.8	10.0	1617.5	104.2	1178.1	9.9
LASSO-BIC	10.0	2307.4	73.3	835.0	NA	10.0	4835.0	97.9	1181.1	NA
				$q = 100$						
Flat	100.0	25000.0	140.4	146.4	5.0	100.0	50000.0	207.9	216.0	10.4
Normal-SS	93.2	372.9	139.9	143.7	7.2	10.0	2675.5	74.4	**91.2**	5.6
MOM-SS	10.0	1286.2	59.1	**70.2**	7.1	10.0	2197.0	75.6	92.8	6.3
ComBat-MLE	100.0	25000.0	70.8	821.1	42.6	100.0	50000.0	123.1	1157.3	14.3
FastBFA	41.5	976.5	84.8	828.2	18.1	65.8	1956.8	130.9	1179.8	13.7
LASSO-BIC	12.3	1663.3	**56.0**	824.7	NA	12.9	3794.4	**70.2**	1167.7	NA



Flat Sparse M | Normal-SS Sparse M | MOM-SS Sparse M | ComBat-SS Sparse M | FastBFA Sparse M | LASSO Sparse M

Figure 26. Scatterplots comparing $ZM^\top$ vs. $\mathbb{E}[Z \mid \widehat{\Delta}, X]\widehat{M}^\top$ between the different models under truly sparse loadings M with $q = 100$ in simulations with batch effect.

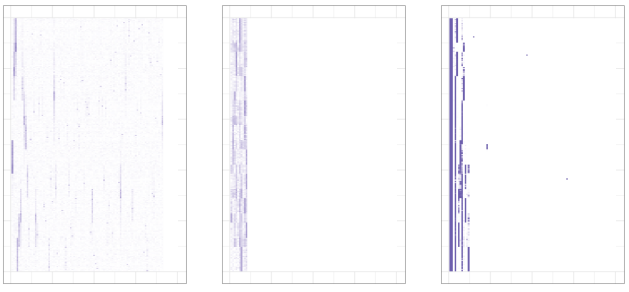
Figure 27 provide a visual representation of the reconstruction of $\widehat{M}, \widehat{\mathrm{Cov}}(\mathrm{x}_i \mid \cdot)^{-1}$ and $\widehat{\gamma}$ for the scenario with truly sparse loadings M, setting $q = 100$ and with mean and variance batch effects.



(A) $\widehat{M}$ Flat (B) $\widehat{M}$ Normal (C) $\widehat{M}$ MOM (D) $\widehat{M}$ ComBat (E) $\widehat{M}$ FastBFA (F) $\widehat{M}$ LASSO

(G) $\widehat{\mathrm{Cov}}(\mathrm{x}_i \mid \cdot)^{-1}$ Flat (H) $\widehat{\mathrm{Cov}}(\mathrm{x}_i \mid \cdot)^{-1}$ Normal (I) $\widehat{\mathrm{Cov}}(\mathrm{x}_i \mid \cdot)^{-1}$ MOM (J) $\widehat{\mathrm{Cov}}(\mathrm{x}_i \mid \cdot)^{-1}$ ComBat (K) $\widehat{\mathrm{Cov}}(\mathrm{x}_i \mid \cdot)^{-1}$ FastBFA (L) $\widehat{\mathrm{Cov}}(\mathrm{x}_i \mid \cdot)^{-1}$ LASSO

Figure 27. Heatmaps of loadings and covariance (red denotes large negative values, blue large positive values, white denotes zero).



(A) $\widehat{\gamma}$ (Normal) (B) $\widehat{\gamma}$ (MOM) (C) $\widehat{\gamma}$ (FastBFA)

Figure 28. Heatmaps of inclusion probability (white denotes 0, dark blue denotes 1).

3.6 Discussion

We have presented a novel model for variance batch effect adjustment via high-dimensional latent factor regression model, and have shown how our model jontly adjusts the data and reduces dimension. We outlined three different prior configurations for the loadings: flat, Normal-SS and a new type of NLP, MOM-SS. Laplace-tailed extensions were discussed and deeper analyses remain as future work. We obtained deterministic optimisations for our model and gave novel EM algorithms to obtain closed-form posterior modes. The use of sparse models increases the quality of our estimations and eases the interpretation. MOM-SS priors encourage parsimony and selective shrinkage and gave good estimation of the latent cardinality in our examples. MOM-SS provided dimension reduction corrected for differences in location and scale due to batches.

4

Applications to cancer data sets

Technological advances in bioinformatics such as high-throughput sequencing, microarrays, mass spectrometry and single cell genomics allow the gathering of vast amounts of biological data, enabling researchers to create models that explain the complex processes and interactions of biological systems (see Bersanelli et al. (2016) for a recent review). Cancer is a prominent example. Large-scale projects such as The Cancer Genome Atlas (TCGA), Cancer Genome Project (CGP) and the International Cancer Genome Consortium (ICGC), as well as many individual laboratories, are generating extensive amounts of biological data (e.g. gene expression, mutation annotation, DNA methylation profiles, copy number changes) and recording other covariates (e.g. gender, tumour stage, medical treatment, and patient history). These projects aim to give a better understanding of the disease and improve prognosis, prevention, and treatment. However, the large and heterogeneous nature of the data make the analyses and interpretations challenging. Furthermore, such data are often generated under different experimental conditions, when new samples are incrementally added to existing samples, or in analyses coming from different projects, laboratories, or platforms. Thus, such a data collecting procedure often produces batch effects (Rhodes et al., 2004), which may lead to incorrect conclusions, unless properly adjusted for (Leek et al., 2010; Goh et al., 2017). In the context of bioinformatics, several approaches have been developed for removing batch effects (see Scherer (2009) for a review and examples). These include data "normalisation" methods that use control metrics or regression methods (Schadt et al., 2001; Yang et al., 2002), matrix factorisation (Alter et al., 2000; Benito et al., 2004), and location-scale methods (Leek and Storey, 2007; Johnson and Li, 2009; Parker et al., 2014; Hornung et al., 2016).

Our examples focus on cancer-related gene expression. We applied our new methods to high-dimensional datasets related to ovarian, lung, and pancreatic cancer.

We compared our methods with the ComBat empirical Bayes batch effect correction

of Johnson et al. (2007), doing an MLE estimation of the factor analysis model after adjusting the data via EM algorithm (ComBat-MLE). We used the **R** package `sva 3.26.0` for ComBat (Leek et al., 2017) and the `fa.em` function in the `cate` package (Wang and Zhao, 2015) for the factor analysis estimation. The **R** code for our model is available at `https://github.com/AleAviP/BFR.BE`.

4.1 Applications to public cancer data sets

We considered two main tasks: giving a visual representation of the latent factors of the data (unsupervised dimension reduction) and carrying out a supervised survival analysis, using the factors obtained in our method as covariates. The aim of the latter is to obtain an external validation of the fact that the extracted factors indeed capture underlying biological signal.

4.1.1 The Clinically Annotated Data for the Ovarian Cancer Transcriptome

According to the World Health Organisation (WHO), ovarian cancer was the seventh most common cancer among women and the eighth leading cause of cancer death worldwide in 2012. Death from ovarian cancer is more common in North America and Europe than in Africa and Asia (World Health Organization, 2014). In the UK, 35% of women who had ovarian cancer survived for ten years or more; in the last forty years the survival rate has almost doubled (NHS, 2015). Carcinomas, five sub-types of which are known, are the most common type of ovarian cancer. The prognosis is generally poor for patients with malignant ovarian tumours, which represent the most lethal gynaecological diseases (World Health Organization, 2014).

A wide range of projects around the world are trying to understand its complex biological origins, features, interactions, and evolution. The "Clinically Annotated Data for the Ovarian Cancer Transcriptome" (Ganzfried et al., 2013) gathers information from several ovarian cancer projects and provides a collection for manually curated gene expression data.

In this thesis we combined information from two datasets of the public **R** package `curatedOvarianData 1.16.0` (Ganzfried et al., 2013). The first was the Illumina Human microRNA array expression dataset `E.MTAB.386`, formed by Angiogenic mRNA and microRNA gene expression signature with $n_1 = 129$ patients (Bentink et al., 2012). The second was the NCI-60 GEO dataset `GSE30161`, consisting of multi-gene expression predictors of single drug responses to adjuvant chemotherapy in ovarian carcinoma for $n_2 = 52$

patients (Ferriss et al., 2012). Prior to our analyses, we selected the 10% genes with highest total variance across all samples obtaining $p = 1,007$ genes. The reason for applying such a pre-selection is that, as recommended by Hackstadt and Hess (2009), microRNAs or genes that show little variance across patients typically carry little biological information. Figure 29 provides a histogram of the gene expression variance across all samples. Subsequently, we normalised all data sets to zero mean and unit variance. We included the age at initial pathologic diagnosis as a covariate.

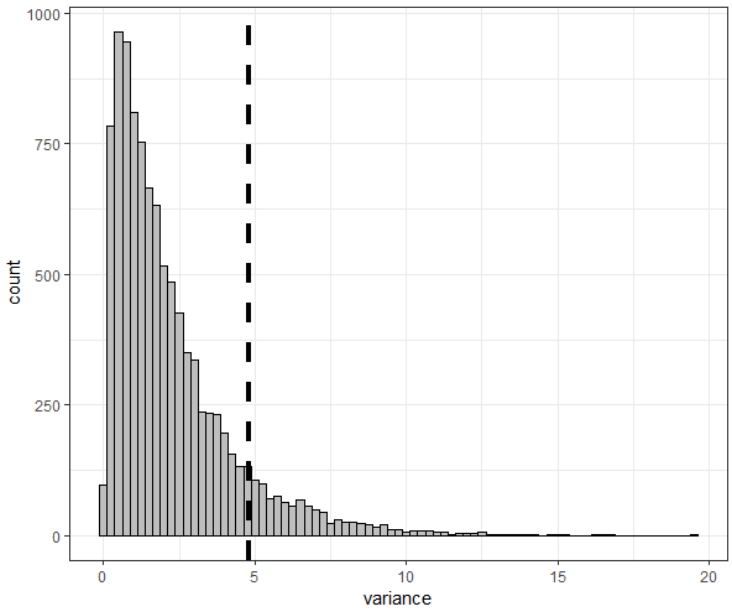
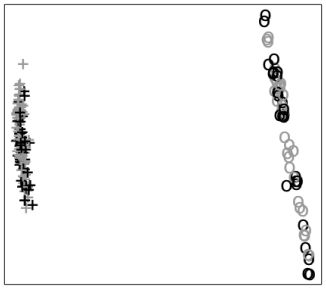


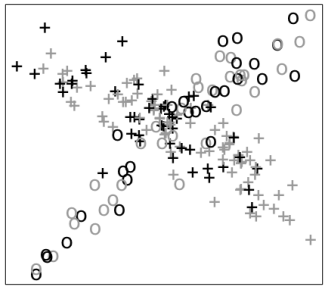
Figure 29. Histogram of the gene expression variance across all samples for `E.MTAB.386` and `GSE30161` ovarian cancer datasets.

Unsupervised: data visualisation

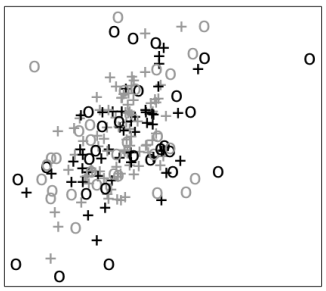
Our first goal was to demonstrate the usefulness of our method as a data visualisation tool. We remark that there are no other model-based approaches that jointly adjust for batch effects and estimate latent factors. Thus, for comparison, we first corrected the data using ComBat and then estimated the parameters via MLE. In order to estimate the number of factors for ComBat-MLE, we carried out a principal component analysis to the corrected ComBat data prior to factor analysis and chose a number of components $\widehat{q}$ that explained 90% or 70% of the total variance.



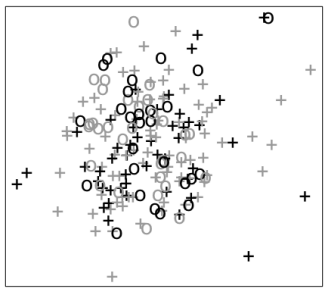
(A) No batch effect correction



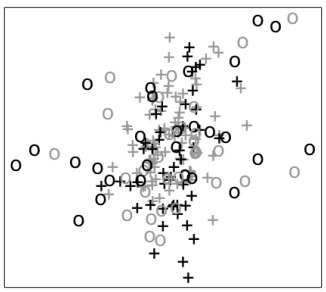
(B) ComBat-MLE



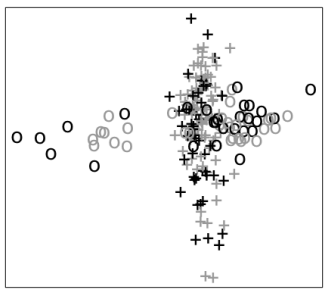
(C) MOM-SS standardised



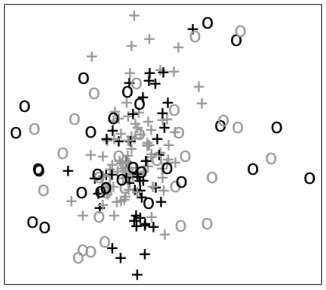
(D) MOM-SS not standardised



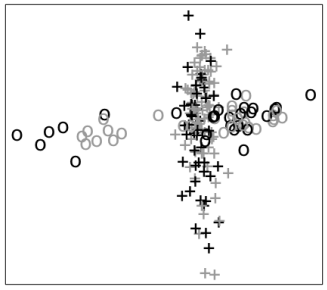
(E) Flat standardised



(F) Flat not standardised



(G) Normal-SS standardised



(H) Normal-SS not standardised

Figure 30. Scatterplot of the first two factors of ovarian cancer dataset for the two different batches (pluses and circles), displaying in black the patients who died within the first three years.

Figure 30 shows the results. We can clearly see the usefulness of ComBat correction (panel (B)) compared to scenario without correction (panel (A)): ComBat removes systematic differences in location and scale across the two batches. Nonetheless, the latent coordinates display distinct covariances. Such covariances are not present in the MOM-SS latent factors, post-processed to standardise their variance $\mathrm{Cov}(\mathsf{z}_i \mid \widehat{\Delta}, X)$ as explained in Section 3.4.4 (panel (C)): one can identify the two factors that contribute the most to the total variance, i.e. with highest $\sum_{j=1}^{p} \widehat{m}_{jk}^2$. Panel (D) displays the same factors as in (C), but without post-processing: this shows the advantages of the standardisation processing step. Finally, we see that Flat and Normal-SS achieved similar performance as MOM-SS. Appendix B displays higher-order latent factors.

Supervised: survival analysis

Here we illustrate the potential of our method as a surpervised tool, performing a survival analysis that aims to predict the time until death. To do this, we applied a Cox proportional hazards model (Cox, 1972), using as covariates the age at initial diagnosis and the latent coordinates obtained in our models. We used the `coxph` function from the **R** package `survival 2.38` (Therneau, 2015) and then assessed the quality of our predictions via the concordance index. Such an index is a non-parametric metric that quantifies the power of a prediction rule via a pair-wise comparison that measures the probability of concordance between the predicted and the observed survival times (Harrell Jr. et al., 1982). To obtain the concordance index we used the function `concordance.index` from the **R** package `survcomp` (Schröeder et al., 2011). The results presented are from 10 independent runs of 10-fold cross-validation. It is important to notice that we did an over-optimistic assessment of ComBat-MLE: we performed a cross-validated factor analysis over the whole ComBat-corrected data, as opposed to adjusting the data via ComBat in an out-of-sample fashion.

Additionally to the other initialisations discussed in Section 3.4.3, we initialised MOM-SS with the values obtained for the Flat prior, and chose the final solution with smallest leave-one-out cross-validated concordance index.

Table 5 shows that MOM-SS achieved a concordance index similar to ComBat-MLE (both 90% and 70%) and a bit higher than Normal-SS, with considerably fewer factors. MOM-SS proved to be a highly competitive sparse model, obtaining a similar predictive performance as the other methods but with 4 latent factors only.

Table 5: Supervised analysis for gene expression of ovarian dataset ($p = 1,007$ genes).

	$\widehat{q}$	$\|\|\widehat{M}\|\|_0$	Concordance index
Flat	100.0	100700.0	0.623
Normal-SS	9.0	7854.6	0.572
MOM-SS	4.0	4028.0	0.599
ComBat-MLE 90%	101.0	101707.0	0.598
ComBat-MLE 70%	41.0	41287.0	0.593

4.1.2 The Cancer Genome Atlas (TCGA): lung cancer

Lung cancer accounts for 13% of all cancers worldwide and is the most common one in men and the third in women worldwide according to the WHO. In 2012 patients had a five year survival rate between 10% and 15%, making this cancer one of the most aggressive. The 5-year survival rate is low (50%) even for patients with the earliest stage I-A, and rapidly declining to 2% for patients with stage IV. It is responsible for 20% of all cancer deaths, and the leading cause of cancer death for men in 87 countries. Due to its resistance to traditional chemotherapy, new genetics and histology based therapies are being developed and studied (World Health Organization, 2014).

The Cancer Genome Atlas (TCGA) aims to gather information from different projects that study several types of cancer, including lung cancer. In this thesis we used microarray data from two different high-throughput platforms: Affymetrix Human Genome `U133A 2.0` Array with $n_1 = 133$ patients and Affymetrix Human Exon `1.0 ST` Array with $n_2 = 112$ patients (Wan et al., 2016). These data are available from the **R** package `TCGA2STAT 1.2` (Wan et al., 2015). We selected the 10% genes with highest total variance accross all samples, obtaining $p = 1,198$ genes – see Figure 31.

Unsupervised: data visualisation

Akin to the ovarian cancer analyses, we first illustrate the effectiveness of our approach as a data visualisation tool. Figure 32 displays the first two factors obtained for the lung cancer data without batch effect correction (panel (A)), with ComBat-MLE correction (panel (B)), and with MOM-SS with ordering and with/without data standardisation of the factors (panels (C) and (D) respectively). ComBat-MLE analyses and MOM-SS post-processing were carried out as in Section 4.1.1. ComBat-MLE and MOM-SS proved to be effective tools for batch effect correction, as both methods corrected the clear batch bias shown in Figure 32(A). See Appendix C for higher-order latent factors.

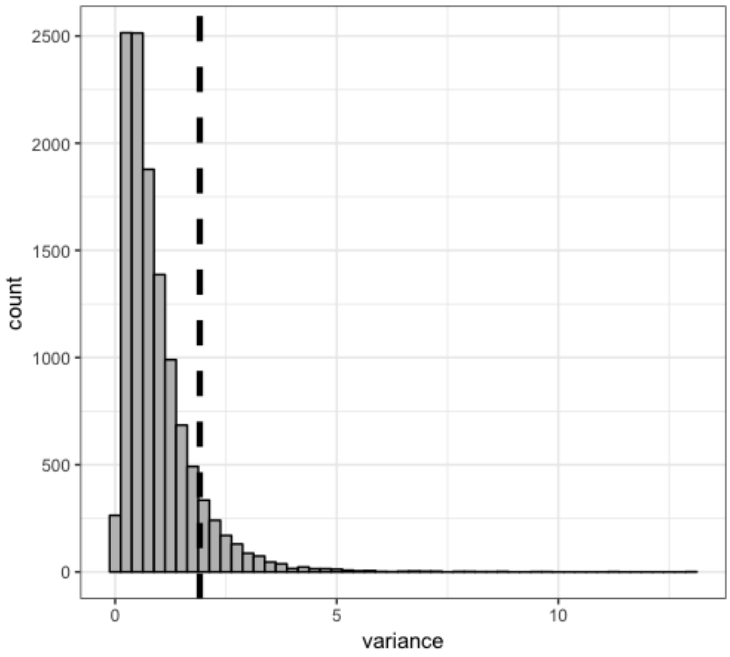


Figure 31. Histogram of the gene expression variance across all samples for the U133A 2.0 and Exon 1.0 ST lung cancer datasets.

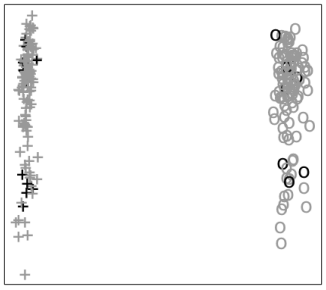
Supervised: survival analysis

We finally show the usefulness of our model for downstream analyses, in particular for survival analysis. As in Section 4.1.1, we performed a Cox proportional hazards model and assessed our method with the concordance index.

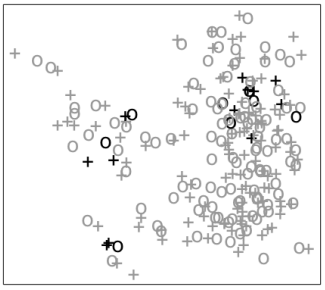
ComBat-MLE 90% and MOM-SS achieved a high concordance index, particularly relative to Normal-SS and ComBat-MLE 70%. Interestingly, MOM-SS recovers a non-sparse solution, resulting in an increased predictive accuracy relative to Normal-SS. Relative to Flat prior, MOM-SS achieves a similar predictive ability with a substantially sparser solution. Despite the good performance of ComBat-MLE, its concordance index proved to be sensitive to the number of factors, being chosen according to either the 70% or 90% rule.

Table 6: Supervised analysis for gene expression of lung cancer dataset ($p = 1,198$ genes).

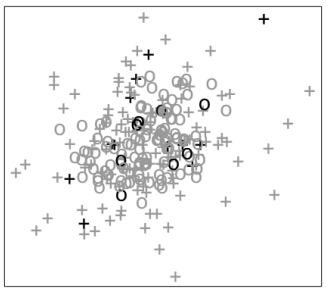
	$\widehat{q}$	$\|\|\widehat{M}\|\|_0$	Concordance index
Flat	100.0	119800.0	0.644
Normal-SS	11.0	13178.0	0.501
MOM-SS	74.0	88652.0	0.640
ComBat-MLE 90%	79.0	94642.0	0.693
ComBat-MLE 70%	30.0	35940.0	0.553



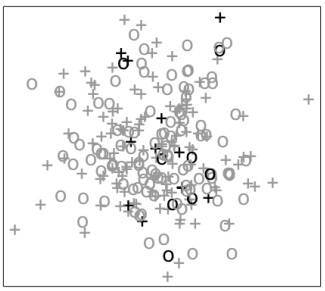
(A) No batch effect correction



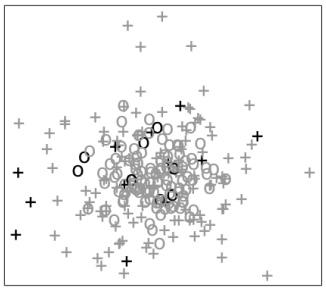
(B) ComBat-MLE



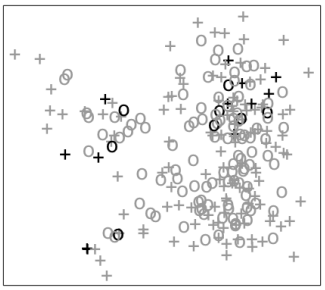
(C) MOM-SS standardised



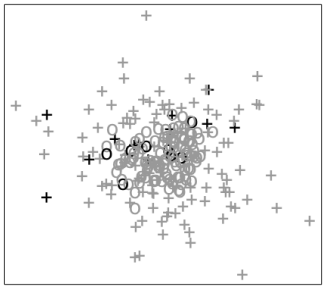
(D) MOM-SS not standardised



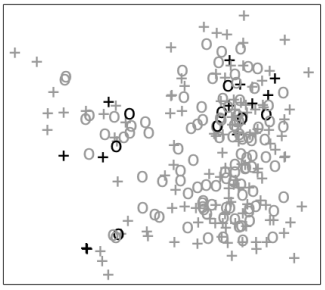
(E) Flat standardised



(F) Flat not standardised



(G) Normal-SS standardised



(H) Normal-SS not standardised

Figure 32. Scatterplot of the first two factors of lung cancer dataset for the two different batches (pluses and circles), displaying in black the patients who died within the first three years.

4.2 Application to off the press Pancreatic Cancer dataset

In 2012, pancreatic cancer was the the 12th most common cancer worldwide according to the WHO, having a particularly high incidence in Europe (with one third of the estimated new cases) and, more in general, in countries at high or very high level of human development. Pancreatic cancer is usually asymptomatic in the early stages and is often not diagnosed until it forms metastases. It has a high fatality rate, with only 5% of the patients surviving 5 years after diagnose. Despite all the efforts to increase the survival rates, progress over the last thirty years has been limited. For more background information, see World Health Organization (2014).

We study an unpublished pancreatic cancer gene expression dataset collected under two different experimental conditions, and different sample size between batches ($n_1 = 27$ and $n_2 = 183$). We assess our approach as a data visualisation tool: our aim is to investigate if MOM-SS has a competitive performance when the sample size significantly differs between batches and when the size of one of the samples is small. We compare our methods with ComBat, which is known to be fairly independent of the sample size in comparison with other batch removal techniques (Johnson et al., 2007). Akin to Sections 4.1.1 and 4.1.2, we estimated the latent cardinality $\widehat{q}$ by selecting the number of components that explained 90% or 70% of the total variance. Subsequently, we selected the 5% genes with the highest total variance across all samples ($p = 1,177$), see Figure 33. We normalised the gene expression to zero mean and unit variance, and post-processed the latent factors as explained in Section 3.4.4 to obtain standardised factors. We included the type of tissue (normal or tumour) as a covariate for our model, i.e. we considered factor regression.

Table 7: Unsupervised analysis for pancreatic cancer dataset ($p = 1,177$ genes).

	$\widehat{q}$	$\|\|\widehat{M}\|\|_0$
Flat	100.0	117700.0
Normal-SS	63.0	74151.0
MOM-SS	19.0	22363.0
ComBat-MLE 90%	97.0	114169.0
ComBat-MLE 70%	32.0	37664.0

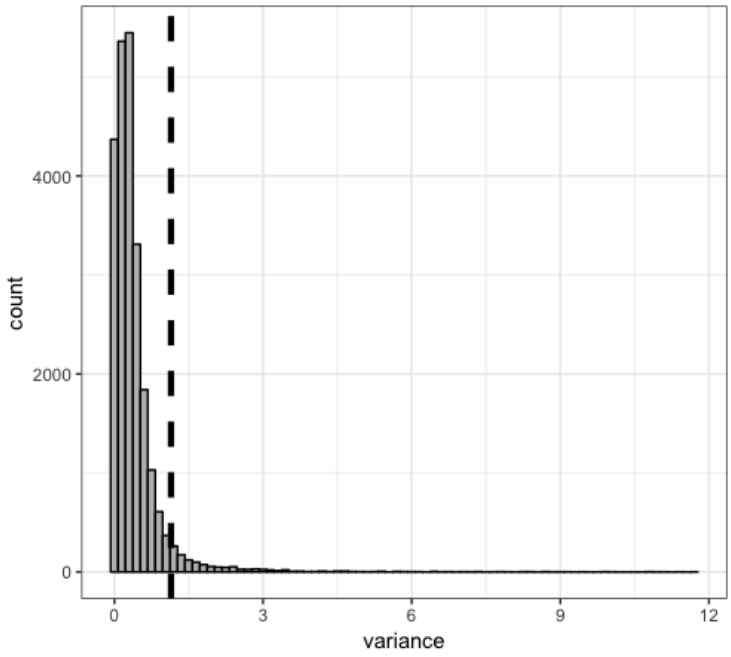
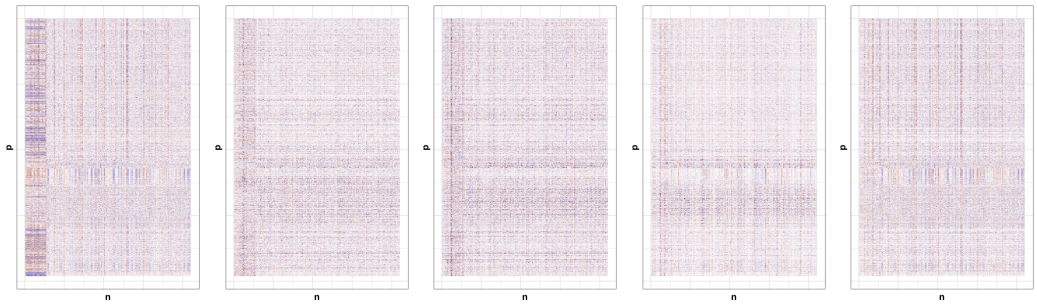
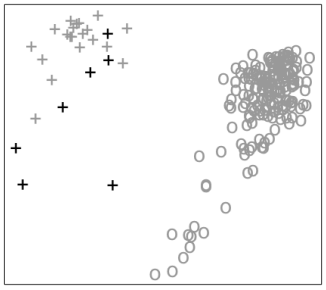


Figure 33. Histogram of the gene expression variance across all samples for pancreatic cancer dataset.

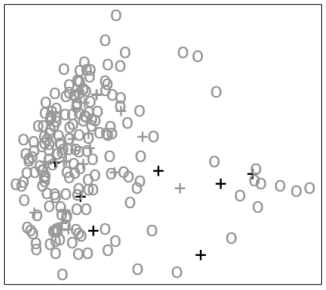


(a) No correction (b) Flat (c) Normal-SS (d) MOM-SS (e) ComBat-MLE

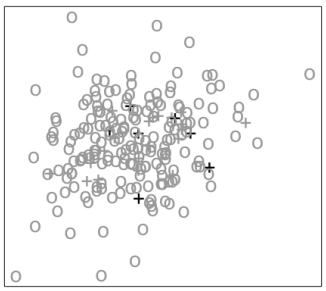
Figure 34. Heatmap of $\mathbb{E}[Z \mid \widehat{\Delta}]\widehat{M}$ for pancreatic cancer dataset (red denotes large negative values, blue large positive values, white denotes zero). This figure shows the advantages of batch effect correction methods.



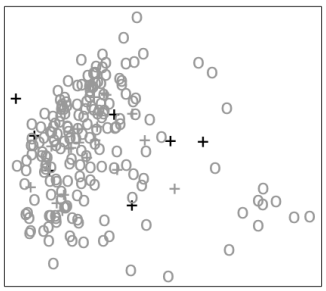
(A) No batch effect correction



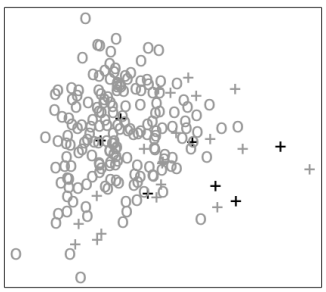
(B) ComBat-MLE



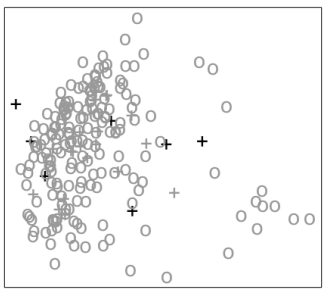
(C) MOM-SS standardised



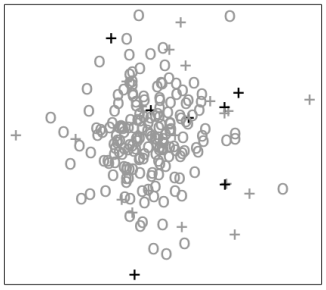
(D) MOM-SS not standardised



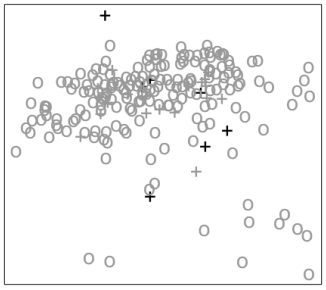
(E) Flat standardised



(F) Flat not standardised



(G) Normal-SS standardised



(H) Normal-SS not standardised

Figure 35. Scatterplot of the first two factors of pancreatic cancer dataset for the two different batches (pluses and circles), displaying in black the patients *without* tumours, i.e. normal.

The tumours from the two different batches are similar in terms of stage, so we would not expect to observe considerably different latent factors. Therefore, the substantial discrepancy in the estimated factor contribution to $\mathbb{E}[X]$, i.e. $\mathbb{E}[Z \mid \widehat{\Delta}]\widehat{M}^\top$, without batch effect correction – see Figure 34(A) – is only due to the different tissue collection and sequencing protocol used. Furthermore, MOM-SS is the model that recovers the sparsest solution for the data, as shown in Table 7. Finally, Figure 35 displays the first two latent factors for the data. Plots for high order factors can be found in Appendix D.

Finally, we asses the quality of our estimators by evaluating the cross-validated log-likelihood, since the underlying data generating truth is unknown. ComBat-MLE does not target the same log-likehood function as our methods, as this is computed using the corrected data (ComBat correction is performed before carrying out an MLE estimation of the FA model); on the contrary, our methods target the original data. We thus assess the out-of-sample predictability comparing our three proposed methods only. Table 8 shows that MOM-SS obtained a better out-of-sample log-likelihood with fewer factors than Normal-SS and Flat. We conclude that MOM-SS reconstructed $\mathrm{Cov}[\mathrm{x}_i]$ better than the other methods.

Table 8: Cross-validated log-likelihood analysis for pancreatic cancer dataset ($p = 1,177$ genes).

	$\widehat{q}$	Log-likelihood
Flat	100.0	-1644.8
Normal-SS	63.0	-1622.0
MOM-SS	19.0	-1157.6

4.3 Discussion

We assessed the quality of our estimations on three different gene expression cancer datasets: ovarian, lung and pancreatic. MOM-SS provided a dimension reduction corrected for mean and variance batch effects, removing distinct covariance patterns present in ComBat-MLE. Furthermore, MOM-SS gave a good visual representation of the data, even when the sample size differs between batches. Finally, MOM-SS achieved a good performance in survival analysis as well as in the evaluation of cross-validated log-likelihoods, often with sparser solutions than other methods.

5

Extensions and future work

In this thesis we presented a model that assumes common factors across the datasets to be integrated. Here we present two direct extensions of this model for future research:

(i) more complex settings where some of the factors differ across batches;

(ii) a model that integrates multiple data types for the same individuals, as opposed to integrating the same variables for different individuals as in Chapter 3.

5.1 Different factors across batches

Our approach can be directly extended to consider a more flexible model that contains batch specific latent variables. This generalisation would allow us to obtain a covariance structure that models the batch specific covariances in addition to the communality, hence it can be of interest in several applications. For instance, in the context of time series, the common factors could reflect the market trends, whereas the batch specific factors may explain recession or expansion periods per batch. Here batch specific factors would be of direct interest to help recover relevant underlying structure. Batch effects may also affect a specific subset of variables, e.g. artefacts may alter a whole image area in a gene expression micro-array study. In this example, batch specific factors would be unwanted structure in the data that needs to be removed in order to perform accurate inference.

Let $\mathsf{x}_i = (x_{i1}, \ldots, x_{ip}) \in \mathbb{R}^p$ be a vector of p observations for the i^{th} individual ($i = 1, \ldots, n$). We model x_i as a regression on observed covariates $\mathsf{v}_i \in \mathbb{R}^{p_v}$, mean batch effects $\mathsf{b}_i \in \mathbb{R}^{p_b}$, latent factors common to all observations $\mathsf{z}_i \in \mathbb{R}^q$ with $q \ll p$, and specific factors per batch $\mathsf{f}_i \in \mathbb{R}^{q_l}$, having different dimension $q_l \ll p$ per $l = 1, \ldots, p_b$ batch. The new Bayesian factor regression model with batch effect correction and specific

factors for individual i in batch l is

$$\mathsf{x}_i = \theta \mathsf{v}_i + \beta \mathsf{b}_i + M\mathsf{z}_i + \Phi^{(l)}\mathsf{f}_i + \mathsf{e}_i, \tag{5.1}$$

where $\theta \in \mathbb{R}^{p \times p_v}$ is the matrix of regression coefficients, $\beta \in \mathbb{R}^{p \times p_b}$ is the matrix of mean batch effects, $M \in \mathbb{R}^{p \times q}$ is the matrix of factor loadings, $\mathsf{e}_i \in \mathbb{R}^p$ models the variance batch effects, and $\Phi^{(l)} = \left\{\Phi^{(l)}_{js}\right\} \in \mathbb{R}^{p \times q_l}$ is the matrix of loadings for batch $l = 1, \ldots, p_b$. We assume: the errors e_{ij} to be $N(0, \tau_{jl}^{-1})$; the common factors z_i to be $N(0, \mathbf{I}_q)$; the specific factors f_i to be $N(0, \mathbf{I}_{q_l})$; all these variables to be independent across all indices and types.

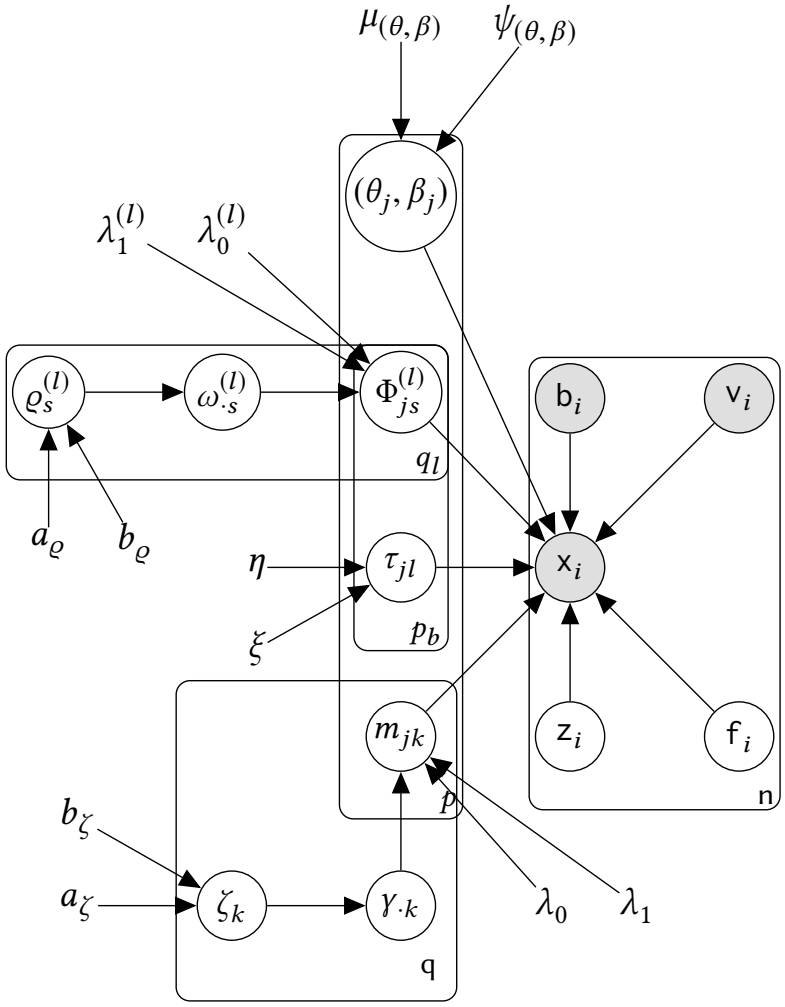


Figure 36. Directed acyclic graph (DAG) for Bayesian Factor Analysis with Batch Effect correction and different factors per batch for Spike-and-slab loadings.

Akin to the common factors, a simple starting strategy would be to set an independent

spike-and-slab (SS) prior on the batch dependent loadings of the form

$$\mathrm{p}(\Phi_{js}^{(l)} \mid \omega_{js}^{(l)}, \lambda_0^{(l)}, \lambda_1^{(l)}) = (1-\omega_{js}^{(l)})\,\mathrm{p}(\Phi_{js}^{(l)} \mid \lambda_0^{(l)}, \omega_{js}^{(l)} = 0) + \omega_{js}^{(l)}\,\mathrm{p}(\Phi_{js}^{(l)} \mid \lambda_1^{(l)}, \omega_{js}^{(l)} = 1), \quad (5.2)$$

for $j = 1, \ldots, p$, $s = 1, \ldots, q_l$, and $l = 1, \ldots, p_b$, where $\mathrm{p}(\Phi_{js}^{(l)} \mid \lambda_0^{(l)}, \omega_{js}^{(l)} = 0)$ is the continuous spike density with dispersion $\lambda_0^{(l)}$ and $\mathrm{p}(\Phi_{js}^{(l)} \mid \lambda_1^{(l)}, \omega_{js}^{(l)} = 1)$ the slab density with dispersion $\lambda_1^{(l)} > \lambda_0^{(l)}$. As before, we propose the following hierarchical prior over $\Phi_{js}^{(l)}$:

$$\begin{aligned} \Phi_{js}^{(l)} \mid \varrho_s^{(l)} &\sim \text{Bernoulli}(\varrho_s^{(l)}), \\ \varrho_s^{(l)} \mid a_\varrho, b_\varrho &\sim \text{Beta}\left(\frac{a_\varrho}{s}, b_\varrho\right), \end{aligned} \quad (5.3)$$

with independent $\Phi_{js}^{(l)}$ for $j = 1, \ldots, p$ and $s = 1, \ldots, q_l$, and default values $a_\varrho = b_\varrho = 1$. Figure 36 provides the DAG for this model.

A possible extension would be to allow for dependence, e.g. a variable with non-zero loadings in batch one could be assigned higher prior probability of non-zero loadings in batch two. This would permit modelling the relationships and interactions between batches. Each factor would then describe dependencies between some of the feature groups. For instance, one might use this approach to model the interactions between brain regions of interest, by calculating correlations between MRI blood-oxygen-levels signals of different regions Klami et al. (2015).

5.2 Integrative model-based factor analysis

Our model can be further extended by combining multiple sets of measured variables per individual. This new setting is of interest when aggregating different data sources for the same individual. For instance, in biomedical applications one may wish to integrate the standard medical history of a patient with other types of data such as demographic, gene expression, RNA-sequencing data, etc. We then extend our model by adding different latent factors per data source. In this setting, the common factors capture the shared covariance structure across all variables and the specific factors model dependencies for each data source. Figure 37 presents a visual representation of this approach, when adding three different datasets with p_1, p_2, and p_3 variables each respectively.

Consider S multiple data types $\mathrm{x}_i^{(s)} \in \mathbb{R}^{p_s}$, $s = 1, \ldots, S$. Let

$$\mathrm{x}_i^\top = \left((\mathrm{x}_i^{(1)})^\top, (\mathrm{x}_i^{(2)})^\top, \ldots, (\mathrm{x}_i^{(S)})^\top\right) \in \mathbb{R}^p,$$

with $p = p_1 + p_2 + \cdots + p_S$. The joint data are modelled as a regression of the common factors

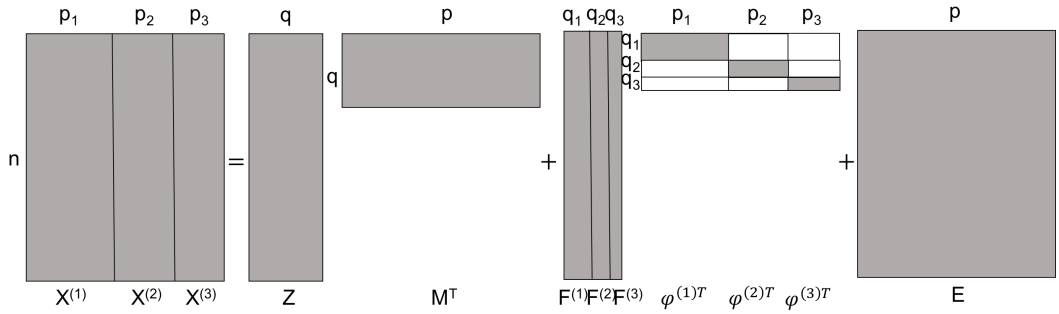


Figure 37. Group factor analysis.

$z_i \in \mathbb{R}^q$ and data source specific factors $f_i^{(s)} \in \mathbb{R}^{q_s}$, $q_s \ll p$, with $f_i = \left(f_i^{(1)}, f_i^{(2)}, \ldots, f_i^{(S)}\right)$. More precisely, we set

$$x_i = Mz_i + \varphi f_i + e_i \tag{5.4}$$

where $\varphi = \left(\varphi^{(1)}, \ldots, \varphi^{(S)}\right)$, with $\varphi^{(s)} \in \mathbb{R}^{p \times p_s}$ for $s = 1, \ldots, S$, is the matrix of data specific loadings, $M \in \mathbb{R}^{p \times q}$ is the matrix of factor loadings, and $e_i \in \mathbb{R}^p$ models the variance. We assume: the errors e_i to be $N(0, \mathcal{T}^{-1})$; the common factors z_i to be $N(0, \mathbf{I}_q)$; the specific factors $f_i^{(s)}$ to be $N(0, \mathbf{I}_{q_s})$; all these variables to be independent across all indices and types.

The resulting covariance of x_i is of the form

$$\Sigma = MM^\top + \varphi\varphi^\top + \mathcal{T}^{-1}. \tag{5.5}$$

This can be decomposed into three different elements:

(i) a low rank approximation of the covariance of the common factors $MM^\top$,

(ii) a block diagonal covariance matrix for the data specific factors $\varphi\varphi^\top$, and

(iii) a diagonal matrix for the common errors $\mathcal{T}^{-1}$.

Moreover, we can easily extend 5.4 to a factor regression model. Now x_i is defined as a regression of the latent factors (z_i and f_i) and other observed covariates $v_i \in \mathbb{R}^{p_v}$:

$$x_i = \theta v_i + Mz_i + \varphi f_i + e_i, \tag{5.6}$$

where $\theta \in \mathbb{R}^{p \times p_v}$ is the matrix of regression coefficients.

In addition to the priors specified in Chapter 3, we set the following priors ($j = 1, \ldots, p$, $k = 1, \ldots, q$, $m = 1, \ldots, q_S$):

- for batch specific loadings $\varphi_{jm}^{(s)}$,

$$\mathrm{p}(\varphi_{jm}^{(s)} \mid \lambda_0^{(s)}, \lambda_1^{(s)}, \omega_{jm}^{(s)}) = (1 - \omega_{jm}^{(s)})\, \mathrm{p}(\varphi_{jm}^{(s)} \mid \lambda_0^{(s)} \omega_{jm}^{(s)} = 0) + \omega_{jm}^{(s)}\, \mathrm{p}(\varphi_{jm}^{(s)} \mid \lambda_1 \gamma_{jm} = 1);$$

- for batch specific latent indicators ω_{jm},

$$\begin{aligned} \omega_{jm} \mid \varrho_s &\sim \text{Bernoulli}(\varrho_s), \\ \varrho_s \mid a_\varrho, b_\varrho &\sim \text{Beta}\left(\frac{a_\varrho}{s}, b_\varrho\right). \end{aligned}$$

Figure 38 provides the DAG for this model.

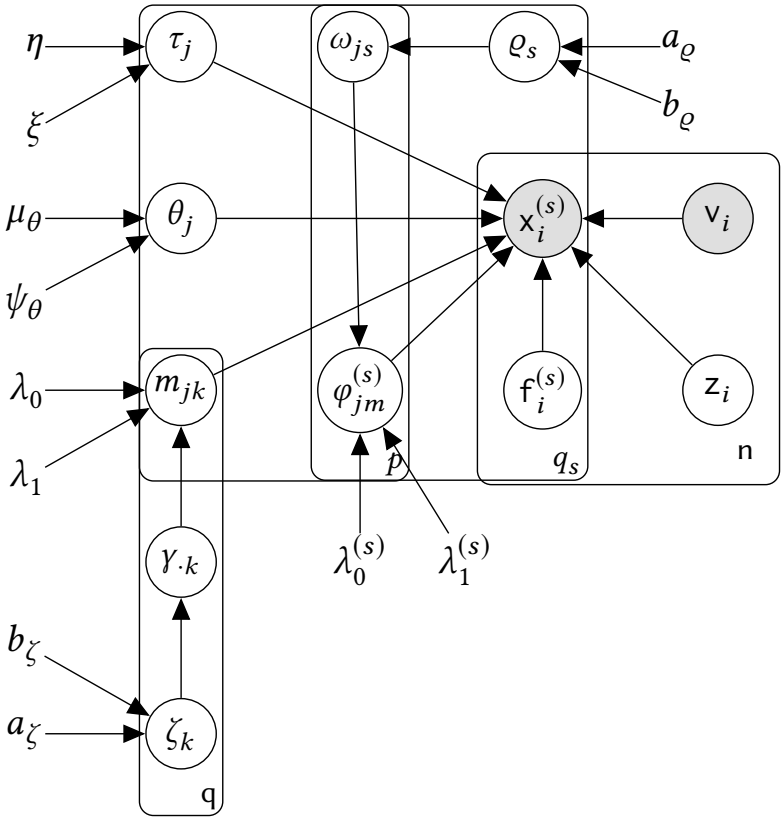


Figure 38. Directed acyclic graph (DAG) for Bayesian factor regression with specific factors per data source and Spike-and-slab prior for common and specific loadings.

We now present some preliminary results for the model in (5.6). We simulated data with $n = 200$, $p_1 = 100$, $p_2 = 100$, $q^* = 10$, $q_1^* = 3$, $q_2^* = 2$, $p_v = 1$, and considered the dense loadings M^* in Figure 14 and the dense batch specific loadings $\varphi^{(s)}$, $s = 1, 2$, in Figure 40(e). We sampled latent factors z_i from $N(0, \mathbf{I}_q)$, errors e_i from $N(0, \mathbf{I}_q)$, covariates

v_i from a continuous Uniform$(0, 3)$, and batch specific latent factors $f_i^{(s)}$ from $N(0, \mathbf{I}_{q_s})$ for $s = 1, 2$. We set the first $p/4$ values of $\theta^* \in \mathbb{R}^p$ to -2 and the other $3p/4$ to 2. Table 9 displays: the selected number of common factors $\widehat{q}$ and batch specific factors $\widehat{q}_s$; the number of estimated non-zero common loadings $||\widehat{M}||_0$, and non-zero batch specific loadings $||\widehat{\varphi}||_0$; the Frobenius norm (F.N.) of the difference between the true expected value and its reconstruction

$$||E[X] - \widehat{E}[X]||_F = ||(ZM^\top + V\theta^\top + F\varphi^\top) - (\mathbb{E}[Z \mid \widehat{\Delta}, X]\widehat{M}^\top + V\widehat{\theta}^\top + \mathbb{E}[F \mid \widehat{\Delta}, X]\widehat{\varphi}^\top||_F,$$

between the true and reconstructed loadings $||ZM^\top - \mathbb{E}[Z \mid \widehat{\Delta}, X]\widehat{M}^\top||_F$, and between the true and reconstructed covariances

$$||\text{Cov}[x_i] - \widehat{\text{Cov}}[x_i]||_F = ||(MM^\top + \varphi\varphi^\top + \mathcal{T}^{-1}) - (\widehat{M}\widehat{M}^\top + \widehat{\varphi}\widehat{\varphi}^\top + \widehat{\mathcal{T}}^{-1})||_F;$$

the number of iterations until convergence.

Table 9: Synthetic data with batch effects and batch specific factors for $n = 200$, $q^* = 5$, $q_1^* = 3$, $q_2^* = 2$, $p_1 = 100$, $p_2 = 100$ parameters, dense loadings M^*, φ^*.

	$\widehat{q}$	$\|\|\widehat{M}\|\|_0$	$\widehat{q}_S$	$\|\|\widehat{\varphi}\|\|_0$	$\|\|\mathbb{E}[X] - \widehat{\mathbb{E}}[X]\|\|_F$	$\|\|ZM^\top - \mathbb{E}[Z \mid \cdot]\widehat{M}^\top\|\|_F$	$\|\|\text{Cov}[x_i] - \widehat{\text{Cov}}[x_i]\|\|_F$	it
	$q = 5, q_s = \{3, 2\}$							
Flat	5	1000	{3,2}	1000	88.52	188.57	108.92	4
Normal-SS	5	839	{3,2}	212	86.27	184.76	83.39	5
MOM-SS	5	690	{3,2}	199	86.21	184.74	79.94	6
	$q = 100, q_s = \{10, 10\}$							
Flat	100	20000	{10,10}	4000	169.85	254.34	137.42	5
Normal-SS	12	1384	{0,0}	0	73.65	203.82	107.02	7
MOM-SS	10	1135	{0,0}	0	67.92	203.76	108.43	9

Table 9 shows that MOM-SS and Normal-SS obtained a good performance relative to Flat prior, assuming a correct guess for the number of common latent factors $q = q^* = 5$ and the batch specific latent factors $q_s = q_s^* = \{3, 2\}$; MOM-SS was overall the model with smallest F.N.'s and highest sparsity. We now illustrate our model in the case $q = 100$ and $q_s = \{10, 10\}$. MOM-SS estimated correctly the total number of latent factors $q + q_1 + q_2 = 10$; however all the covariance is modelled by the common factors, as the value returned for the batch specific cardinality is zero. Deeper analysis of this phenomenon remains as future work. In the same setting, MOM-SS and Normal-SS achieved smaller F.N.'s than Flat, reflecting the advantages of shrinkage.

Figure 39 shows the log-likelihood of the three methods when setting $q = 100$. MOM-SS provides the highest log-likelihood.

Finally, Figure 40 provides a visual representation of the reconstructions of $\widehat{M}$, $\widehat{\varphi}$, $\widehat{\mathcal{T}}^{-1}$,

and $\widehat{\mathbb{E}}[X]$ when setting $q = 5, q_s = \{3, 2\}$. These plots display a good reconstruction of the underlying data-generating loadings, errors, etc. Further analysis of the Bayesian factor regression with specific factors per data source model is yet to be done.

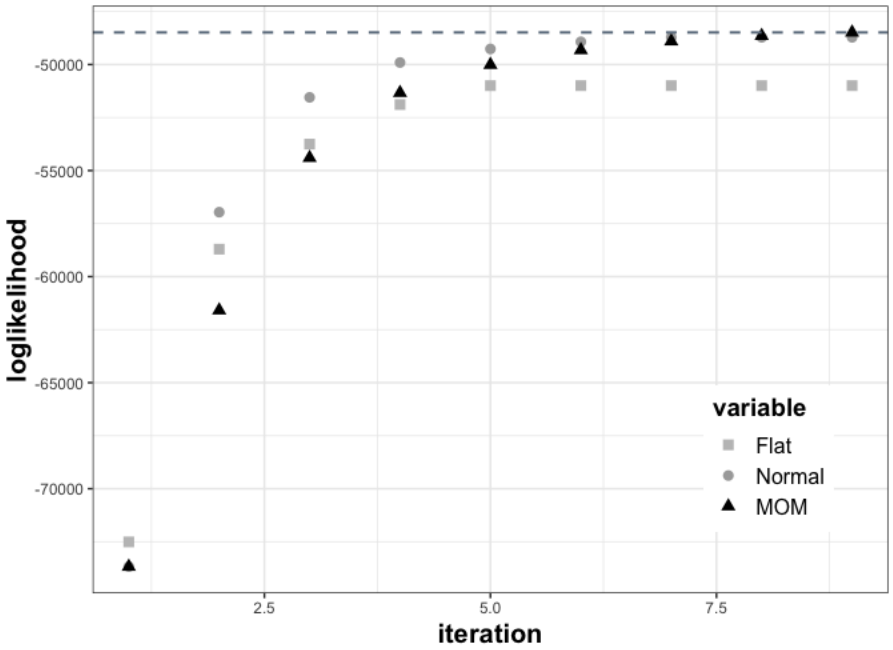


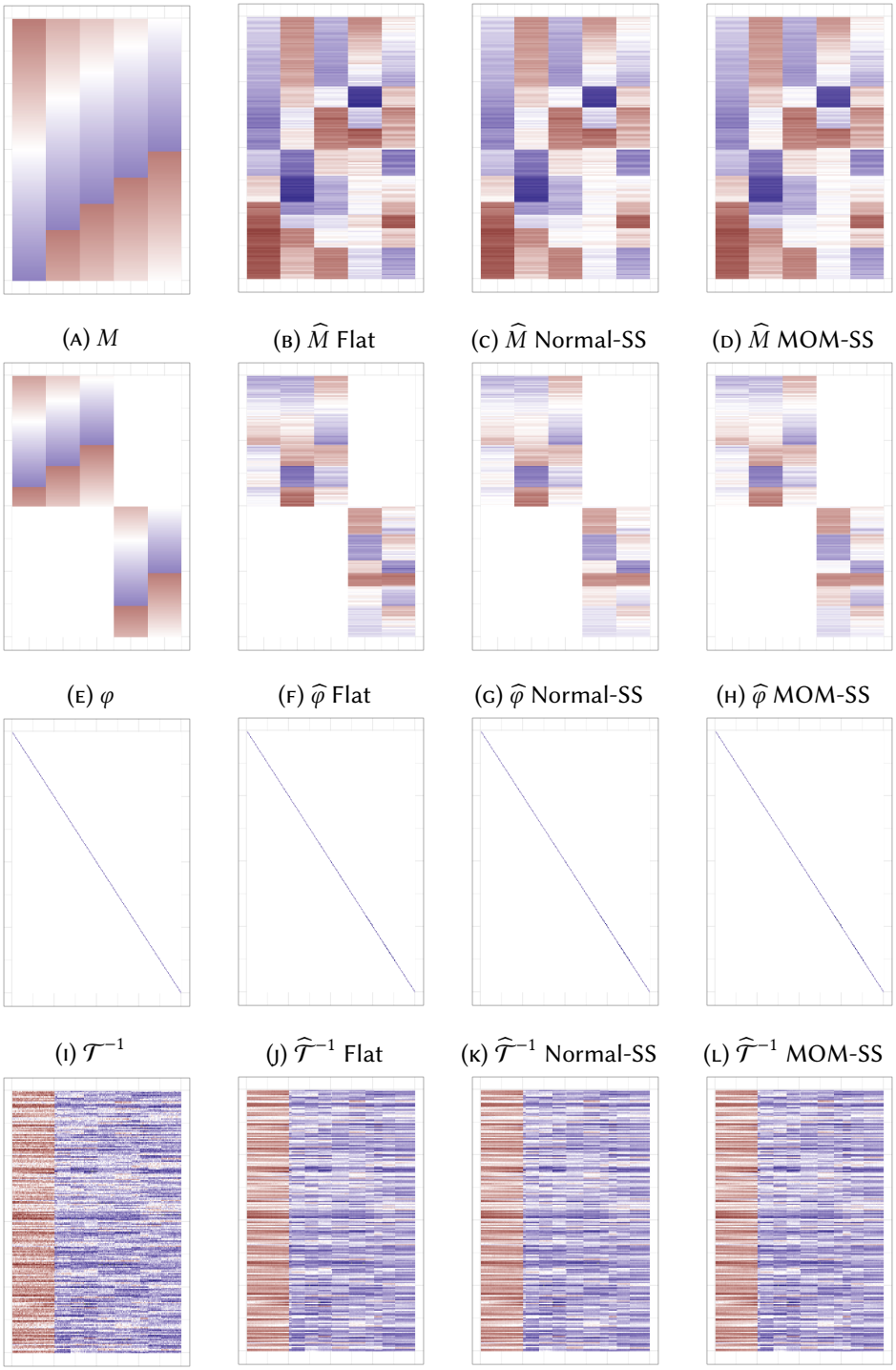
Figure 39. Comparison of the log-likelihood convergence for Flat, Normal-SS, and MOM-SS priors for Bayesian factor regression with specific factors per data source model. Log-likelihood value at convergence for MOM-SS in dotted line.

5.3 Further extensions

Some further extensions not discussed here include scenarios where the same variables are only recorded for a subset of the individuals or where different types of datasets, such as continuous (e.g. methylation profiles), count (e.g. RNA-sequencing and gene expression), zero-inflated (e.g. Single-cell), and binary (e.g. mutation) data are integrated.

Throughout this thesis, non-local-based spike-and-slab priors have been applied to factor analysis. As a possible extension these priors could be studied in different settings, such as generalised linear models or graphical models.

Finally, methods to assess the uncertainty in our point estimates could be developed via MCMC algorithms or variational inference.



(A) M (B) $\widehat{M}$ Flat (C) $\widehat{M}$ Normal-SS (D) $\widehat{M}$ MOM-SS

(E) φ (F) $\widehat{\varphi}$ Flat (G) $\widehat{\varphi}$ Normal-SS (H) $\widehat{\varphi}$ MOM-SS

(I) $\mathcal{T}^{-1}$ (J) $\widehat{\mathcal{T}}^{-1}$ Flat (K) $\widehat{\mathcal{T}}^{-1}$ Normal-SS (L) $\widehat{\mathcal{T}}^{-1}$ MOM-SS

(M) X (N) $\widehat{\mathbb{E}}[X]$ Flat (O) $\widehat{\mathbb{E}}[X]$ Normal-SS (P) $\widehat{\mathbb{E}}[X]$ MOM-SS

Figure 40. Heatmaps for Bayesian factor regression with specific factors per data source model (white denotes 0, dark blue denotes positive values, and red denotes negative values).

6

Discussion

In this thesis we have presented a novel model for latent factor regression and variance batch effect adjustment, and have shown how to jointly adjust the data and reduce dimension, and obtain sparse covariance estimates. We outlined three different prior configurations for the loadings: flat, Normal-spike-and-slab (Normal-SS), and a new type of Non-local priors (NLPs), i.e. Normal-spike-MOM-slab (MOM-SS). We discussed Laplace-tailed extensions, but deeper analyses remain as future work. To our knowledge this is the first time NLPs are implemented in the factor analysis context. We gave deterministic optimisations for our model and provided novel EM algorithms to obtain closed-form posterior modes. We showed that the use of sparse models increases the quality of parameter estimations, even in the absence of batches. MOM-SS priors proved to be appealing, improving the estimation of factor cardinality and encouraging parsimony and selective shrinkage.

We illustrated the usefulness of our method in unsupervised and supervised applied data analysis. MOM-SS provided dimension reduction corrected for differences in location and scale due to batches, removing distinct covariances' patterns present in two-stage methods that adjust variances and fit the factor model separately. Our model demonstrated to be useful for downstream analyses when performing survival analysis for cancer patients (i.e. predicting likely survival times) achieving a competitive concordance indexes, in some cases with substantially fewer factors. Although we focused on gene expression of cancer datasets, we remark that our model-based approach could be usefully applied in other settings.

In this thesis we obtained only point estimates of the posterior mode, which are the main object of interest in dimensionality reduction and in the supervised survival prediction problems we considered here. However, in situations where one wishes to perform inference on the covariance, it would be useful to assess the uncertainty in the point es-

timates. A natural strategy for this would be to use MCMC algorithms, or alternatively some faster-but-inexact alternatives such as variational inference. Such an uncertainty quantification is left as future work.

It is important to highlight that under Laplace-spike-and-slab (Laplace-SS) one can obtain exact sparsity in the factor loadings, but this is not the case for the other prior configurations outlined in this thesis (flat, Normal-SS, MOM-SS and Laplace-MOM-SS). Thus, a post-processing step is required to effectively select the number of factors and non-zero loadings. Therefore, the EM algorithms proposed in this thesis can be seen as regularisation methods more than selection techniques. A deeper analysis of our models under Laplace-SS as a selection method remains as future research.

We also remark that our novel MOM-SS and its closed-form EM updates can be extended to different interesting frameworks beyond factor models, such as linear or multivariate regression models, generalised linear models, and graphical models.

It should be emphasised that our model assumes common factors across the datasets being integrated. Thus, as a direct extension for future research, one may consider more complex settings, where e.g.: some of the factors differ across data sources; one needs to add different variables from the various datasets to be integrated (as opposed to adding different individuals as in our case); the same variables are possibly recorded for a subset of the individuals only. Some of these extensions were discussed in Chapter 5.

In this era, many disciplines are able to generate increasingly more and larger volumes of data. To work effectively with these, it is crucial to have strong statistical tools that address the important challenges associated. In this thesis, we have presented a new class of models that achieve an integrated visualisation and improved interpretation of data coming from different studies.

Appendix

A Reproducibility: R package

A.1 Getting started

Our **R** package can be installed via devtools:

```
install.packages("devtools") #Install devtools
library(devtools)
devtools::install_github("AleAviP/BFR.BE")
library(BFR.BE)
```

It has two main functions: BFR.BE.EM and BFR.BE.EM.CV.

BFR.BE.EM performs our EM algorithm for Bayesian Fator Analysis and Bayesian Factor Regression, with and without batch effect correction for our three prior specifications: MOM-SS (N.MOM), Normal-SS (Normal) and Flat (Flat).

BFR.BE.EM.CV selects the best model from the four different alternatives (two different initialisations and two different post-processing options) outlined in Sections 2.11, 2.12, 3.4.3, and 3.4.4 via our proposed weighted 10-fold cross-validation.

Documentation for both functions is available at:

```
?BFR.BE.EM
?BFR.BE.EM.CV
```

A.2 Bayesian factor analysis

We illustrate the usage of our package for Bayesian factor analysis with the simulated data used in Section 2.13.2 with truly sparse loadings.

We simulate the data as follows:

```
#####Running simulations####
library(BFR.BE) #My code
```

```
set.seed(123)
n<- 100
p<-1000
q<-10

length=trunc((p-1+q)/(.75*q+.25))
offset=trunc(length*.25)
p<-length+(length-offset-1)*(q-1) #New p

M<-matrix(0,p,q)
end<-1
for(i in (1:q)){
  start<-end-(i!=1)*offset
  end<-start+length-1
  M[start:end,i]<-1
}

E <- rmvnorm(n,numeric(p),diag(p))
Z <- rmvnorm(n,numeric(q),diag(q))
X <- Z%*%t(M) + E
Sigma<-M%*%t(M)+diag(p)
```

The EM estimation of the model can be easily done using the **R** function `BFR.BE.EM.CV`, setting `varianceBE=FALSE` (no batch effect correction), and specifying the q^* factors and the prior to use: `N.MOM`, `Normal` or `Flat`. If the user does not select a prior specification for the loadings, the default option is `N.MOM`. Default hyper-parameters are set as in Section 2.10, but the user may specify other parameters that reflect a priori knowledge.

```
##Estimation
#MOM
MOM_10=BFR.BE.EM.CV(x=X,q = 10,varianceBE=FALSE)
MOM_100=BFR.BE.EM.CV(x=X,q = 100,varianceBE=FALSE)

#Normal
Normal_10=BFR.BE.EM.CV(x=X,q = 10,
    prior="Normal",varianceBE=FALSE)
Normal_100=BFR.BE.EM.CV(x=X,q = 100,
    prior="Normal",varianceBE=FALSE)

#Flat
Flat_10=BFR.BE.EM.CV(x=X,q = 10, prior="Flat",varianceBE=FALSE)
Flat_100=BFR.BE.EM.CV(x=X,q = 100, prior="Flat",varianceBE=FALSE)
```

The function returns the EM updates for the $\widehat{\Delta} = (\widehat{M}, \widehat{\mathcal{T}}, \widehat{\zeta})$, the expected values for the latent factors $\mathbb{E}[z_i \mid \widehat{\delta}, X]$, and the latent indicators $\mathbb{E}[\gamma_{jk} \mid \widehat{\delta}]$. It also outputs the post-processed loadings `Mpost` with the smallest weighted 10-fold cross-validated reconstruction error, as in Section 2.12.

A.3 Bayesian factor regression

We now present our Bayesian Factor regression model with batch effect correction in the simulated data of Section 3.5.1 with dense loadings:

```
library(BFR.BE) #My code
#####Simulating the data####
set.seed(123)
n<-200
p<-250
q<-10
Pv = 1
Pb = 2

grid<-seq(-1,1,length.out=p)
M<-matrix(grid,ncol=q,nrow=p)
rate<-trunc(p/(q*2))
for(i in 2:q){
  M[,i]<-grid[c((i*rate):p,1:(i*rate-1))]
}

Z <- rmvnorm(n,numeric(q),diag(q))
Psi <- diag(c(rep(.5,p)),p)
V <- matrix(ncol = Pv, nrow = n, runif(n*Pv,0,3))
Theta = matrix(ncol=Pv,c(rep(-2,round(p/2)),rep(2,p-round(p/2))))
batch = round(runif(n,0,1),0)
B = matrix(ncol=Pb, c(batch,ifelse(batch==1,0,1)))
Beta = matrix(ncol=Pb, c(rep(2,p),rep(0,p)))
Tau_inv = cbind(diag(Psi),diag(Psi)*1.5)
Er = matrix(ncol=p, nrow=n)
for (i in 1:n){
  if(i%%10==0){print(i)}
  Er[i,] = mvrnorm(1, rep(0,p), diag(c(Tau_inv %*% B[i,]),p), tol
      = 1e-6, empirical = FALSE, EISPACK = FALSE)
}
```

```
30
31 X <- Z%*%t(M) + V%*%t(Theta) + B%*%t(Beta) + Er
32
33 wsaq  1 º##Estimation
34 #MOM
35 MOM_10=BFR.BE.EM.CV(x=X,v=V,b=B,q = 10)
36 MOM_100=BFR.BE.EM.CV(x=X,v=V,b=B,q = 100)
37
38 #Normal
39 Normal_10=BFR.BE.EM.CV(x=X,v=V,b=B,q = 10, prior="Normal")
40 Normal_100=BFR.BE.EM.CV(x=X,v=V,b=B,q = 100, prior="Normal")
41
42 #Flat
43 Flat_10=BFR.BE.EM.CV(x=X,v=V,b=B,q = 10, prior="Flat")
44 Flat_100=BFR.BE.EM.CV(x=X,v=V,b=B,q = 100, prior="Flat")
```

BFR.BE.EM.CV outputs the estimations $\widehat{\Delta} = (\widehat{M}, \widehat{\theta}, \widehat{\beta}, \widehat{\mathcal{T}}, \widehat{\zeta}), \mathbb{E}[z_i \mid \widehat{\delta}, X], \mathbb{E}[\gamma_{jk} \mid \widehat{\delta}]$, and the selected post-processed loadings, using the weighted 10-fold cross-validation procedure of Section 3.4.4.

A.4 Tables

In order to reproduce the key tables of this thesis, we provide the following code for the Bayesian factor analysis model:

```
1  q.hat.2=data.frame(Flat_10=10,
2                     Normal_10=sum(apply(Normal_10$gamma>0.5,2,sum)!=0),
3                     MOM_10=sum(apply(MOM_10$gamma>0.5,2,sum)!=0),
4                     Flat_100=100,
5                     Normal_100=sum(apply(Normal_100$gamma>0.5,2,sum)!=0),
6                     MOM_100=sum(apply(MOM_100$gamma>0.5,2,sum)!=0))
7
8  m.hat.2=data.frame(Flat_10=1000*10,
9                     Normal_10=sum(Normal_10$gamma>0.5),
10                    MOM_10=sum(MOM_10$gamma>0.5),
11                    Flat_100=1000*100,
12                    Normal_100=sum(Normal_100$gamma>0.5),
13                    MOM_100=sum(MOM_100$gamma>0.5))
14
15 FN.ZM=data.frame(Flat_10=FNorm(Flat_10$Ez,Flat_10$M,Z,M),
16                  Normal_10=FNorm(Normal_10$Ez,Normal_10$Mpost,Z,M),
17                  MOM_10=FNorm(MOM_10$Ez,MOM_10$Mpost,Z,M),
18                  Flat_100=FNorm(Flat_100$Ez,Flat_100$M,Z,M),
19                  Normal_100=FNorm(Normal_100$Ez,Normal_100$Mpost,Z,M),
20                  MOM_100=FNorm(MOM_100$Ez,MOM_100$Mpost,Z,M))
21
```

```
FN.covX=data.frame(Flat_10=FNorm2(Flat_10$sigma,Flat_10$M,Sigma),
                   Normal_10=FNorm2(Normal_10$sigma,Normal_10$Mpost,Sigma),
                   MOM_10=FNorm2(MOM_10$sigma,MOM_10$Mpost,Sigma),
                   Flat_100=FNorm2(Flat_100$sigma,Flat_100$M,Sigma),
                   Normal_100=FNorm2(Normal_100$sigma,Normal_100$Mpost,Sigma),
                   MOM_100=FNorm2(MOM_100$sigma,MOM_100$Mpost,Sigma))

it=data.frame(Flat_10=Flat_10$iterations,
              Normal_10=Normal_10$iterations,
              MOM_10=MOM_10$iterations,
              Flat_100=Flat_100$iterations,
              Normal_100=Normal_100$iterations,
              MOM_100=MOM_100$iterations)

tableFN=rbind.data.frame(q_hat=q.hat.2,
                         Mhat=m.hat.2,
                         FN_ZM2=FN.ZM,
                         FN_covX=FN.covX,
                         it=it)
round(t(tableFN),1)
```

The functions that return the tables of the Bayesian factor regression model use, in addition to the same code as Bayesian FA, the following:

```
#####Table####
FN.x=data.frame(Flat_10=FNormXhat(Flat_10,Z,V,B,Theta,Beta,M,Flat_10$columns),
                Normal_10=FNormXhat(Normal_10,Z,V,B,Theta,Beta,M,Normal_10$columns),
                MOM_10=FNormXhat(MOM_10,Z,V,B,Theta,Beta,M,MOM_10$columns),
                Flat_100=FNormXhat(Flat_100,Z,V,B,Theta,Beta,M,Flat_100$columns),
                Normal_100=FNormXhat(Normal_100,Z,V,B,Theta,Beta,M,Normal_100$columns),
                MOM_100=FNormXhat(MOM_100,Z,V,B,Theta,Beta,M,MOM_100$columns))

tableFN=rbind.data.frame(q_hat=q.hat.2,
                         Mhat=m.hat.2,
                         FN_X=FN.x,
                         FN_ZM2=FN.ZM,
                         it=it)
round(t(tableFN),1)
```

A.5 Plots

BFR.BE package provides three different functions to reproduce the plots displayed in this thesis: `plot.scat` for the scatterplots, `plot.heat` for the loadings' heatmaps and `plot.heat.cov` for the covariance heatmap.

```
#####Plots####
##Scaterplots
plot.scat(Z%*%t(M),Flat_100$Ez%*%t(Flat_100$M))
```

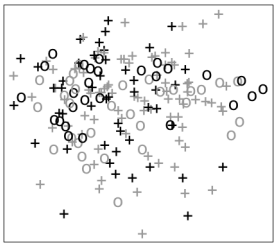
```
plot.scat(Z%*%t(M),Normal_100$Ez%*%t(Normal_100$Mpost))
plot.scat(Z%*%t(M),MOM_100$Ez%*%t(MOM_100$Mpost))

##Heatmaps
##M
#Flat
plot.heat(Flat_100$M,limit=c(-2,2))
#Normal
plot.heat(Normal_100$Mpost,limit=c(-2,2),rotation=TRUE)
#MOM
plot.heat(MOM_100$Mpost,limit=c(-2,2),rotation=TRUE)

##Covariance
#Flat
plot.heat.cov(Flat_100$M,diag(Flat_100$sigma),limit=c(-3,3))
#Normal
plot.heat.cov(Normal_100$M,diag(Normal_100$sigma),limit=c(-3,3))
#MOM
plot.heat.cov(MOM_100$M,diag(MOM_100$sigma),limit=c(-3,3))

##Gamma
plot.heat(Normal_100$gamma)
plot.heat(MOM_100$gamma)
```

A.6 Post-processing of the latent factors

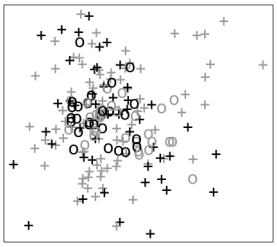
Finally, function FA.ov returns the post-processed standardised factors $\widetilde{z}_i$ for data visualisation (See Section 3.4.4):

```
FA.ov(MOM_100,B)
```

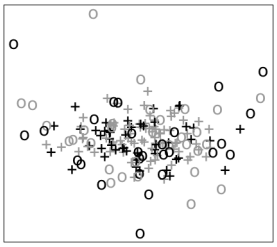
B Ovarian cancer unsupervised: Z_3 vs Z_4 latent factors



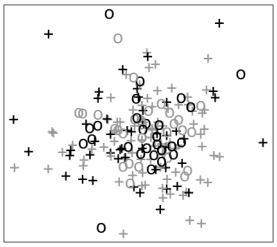
(a) No batch effect correction



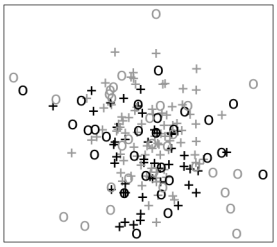
(b) ComBat-MLE



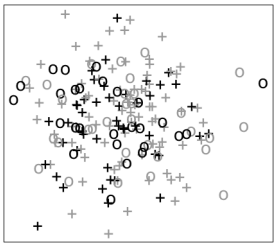
(c) MOM-SS standardised



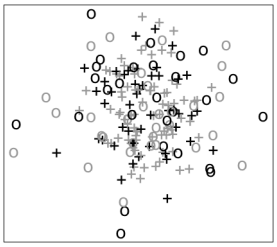
(d) MOM-SS not standardised



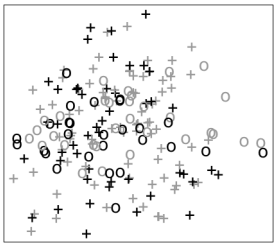
(e) Flat standardised



(f) Flat not standardised



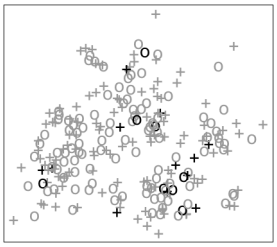
(g) Normal-SS standardised



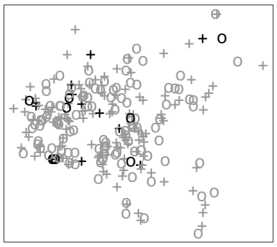
(h) Normal-SS not standardised

Figure 41. Scatterplot of the third and fourth factors of ovarian cancer dataset for the two different batches (pluses and circles), displaying in black the patients who died within the first three years.

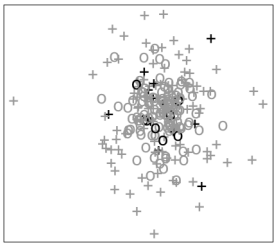
C Lung cancer unsupervised: Z_3 vs Z_4 latent factors



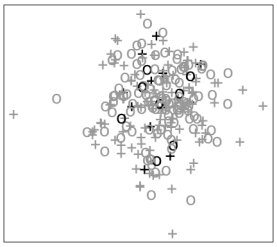
(A) No batch effect correction



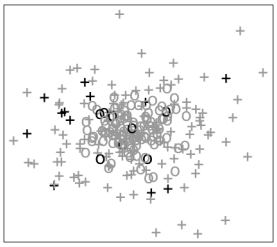
(B) ComBat-MLE



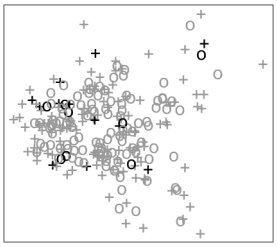
(C) MOM-SS standardised



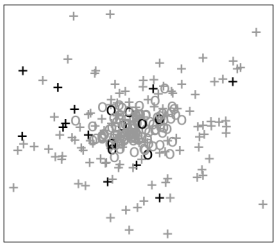
(D) MOM-SS not standardised



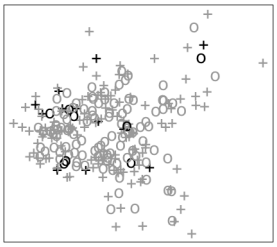
(E) Flat standardised



(F) Flat not standardised



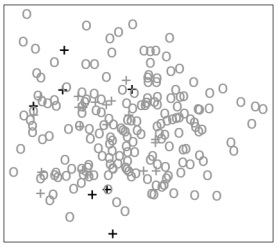
(G) Normal-SS standardised



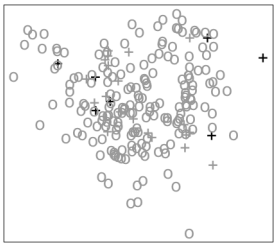
(H) Normal-SS not standardised

Figure 42. Scatterplot of the third and fourth factors of lung cancer dataset for the two different batches (pluses and circles), displaying in black the patients who died within the first three years.

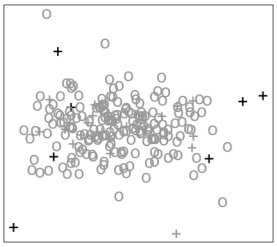
D Pancreatic cancer unsupervised: Z_3 vs Z_4 latent factors



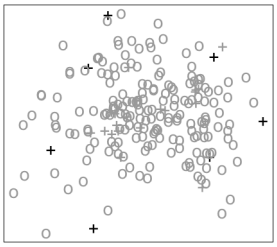
(a) No batch effect correction



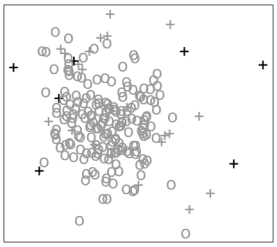
(b) ComBat-MLE



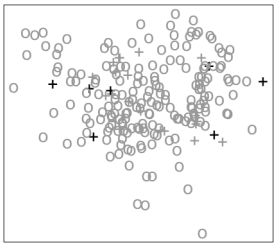
(c) MOM-SS standardised



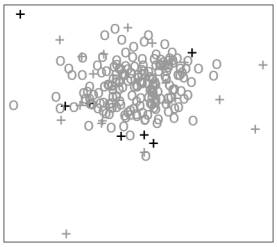
(d) MOM-SS not standardised



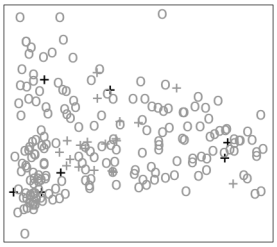
(e) Flat standardised



(f) Flat not standardised



(g) Normal-SS standardised



(h) Normal-SS not standardised

Figure 43. Scatterplot of the first and fourth factors of pancreatic cancer dataset for the two different batches (pluses and circles), displaying in black the patients without tumour.

Glossary

AIC Akaike information criterion.

BECA Batch effect-correction algorithms.

BIC Bayesian information criterion.

BMA Bayesian model averaging.

BMS Bayesian model selection.

CGP Cancer Genome Project.

ComBat Empirical Bayes batch effect correction.

ComBat-MLE ComBat empirical Bayes batch effect correction with an MLE estimation of the factor analysis model.

DAG Directed acyclic graph.

EM Expectation-Maximisation.

eMOM Exponential moment non-local prior.

FA Factor Analysis.

FastBFA Fast Bayesian Factor Analysis via Automatic Rotations to Sparsity.

ICGC International Cancer Genome Consortium.

iMOM Inverse moment non-local prior.

Laplace-MOM-SS Laplace-spike-and-MOM-slab.

LASSO-BIC Penalized Likelihood Factor Analysis with a LASSO penalty.

L-SSs Local spike-and-slabs.

Laplace-SS Laplace-spike-and-slab.

LASSO Least Absolute Shrinkage and Selection Operator.

LP Local priors.

LS Localisation-scale.

MAP Maximum A Posteriori.

MOM Moment non-local prior.

MOM-SS Normal-spike-and-MOM-slab.

NLP Non-local Prior.

Normal-SS Normal-spike-and-slab.

PCA Principal Component Analysis.

PPCA Probabilistic Principal Component Analysis.

SS Spike-and-slab.

SVD Singular value decomposition.

TCGA The Cancer Genome Atlas.

WHO World Health Organisation.

Notation Glossary

$\widehat{a}$	estimatimation of variable a
i	individuals $i = 1, \ldots, n$
j	parameters $j = 1, \ldots, p$
k	low dimensional parameters $k = 1, \ldots, q$
l	bathes $l = 1, \ldots, p_b$
n	number of individuals
p	number of variables
q	latent factor cardinality
t	iterations
l_k	eigen-values
m_j	vector of loadings
n_l	number of individuals in batch l
p_v	number of observed covariates
p_b	number of batches
u_k	eigen-vectors
b_i	indicator vector
e_i	vector of errors
v_i	vector of observed covariates
x_i	vector of observations
z_i	vector of latent coordinates or latent factors
b_{il}	batch indicator $b_{il} := 1$ if individual i is in batch l, $b_{il} := 0$ otherwise.
e_{ij}	error
m_{jk}	loading
$\widehat{p}_{jk}$	$\mathbb{E}[\gamma_{jk} \mid \widehat{\Delta}]$
x_{ij}	observation
z_{ik}	latent coordinate or latent factor

E	matrix of errors
M	matrix of factor loadings
S	sample covariance matrix
T	maximum number of iterations
X	matrix of observations
Z	matrix of latent coordinates or factors
β	mean or additive batch effects
ε	tolerance in the log-posterior change
ε_M	tolerance in the loadings change
Δ	matrix of variables to maximise
θ	matrix of regression coefficients
λ_0	dispersion parameter of the spike component for local priors
λ_1	dispersion parameter of the slab component for local priors
$\widetilde{\lambda}_0$	dispersion parameter of the spike component for non-local priors
$\widetilde{\lambda}_1$	dispersion parameter of the slab component for non-local priors
γ_{jk}	latent indicators
$\gamma_{\cdot k}$	column k of matrix γ
γ	matrix of latent indicators
μ	mean
τ_{jl}	j^{th} idiosyncratic precision element in batch l
$\mathcal{T}_{\mathrm{b}_i}$	idiosyncratic precision matrix in batch l
$\mathcal{T}_l$	idiosyncratic precision matrix for a fixed batch l
$\mathcal{T}^{-1}$	idiosyncratic variances
ζ_k	weights of the latent indicators
$\mathrm{Cov}[\cdot]$	covariance
$\mathbb{E}[\cdot]$	expected value
$Q(\cdot)$	Expected complete-data log-posterior
$\|\| \cdot \|\|$	norm
$\|\| \cdot \|\|_0$	L0 norm
$\|\| \cdot \|\|_F$	Frobenius norm
$A \circ B$	Hadamard (element-wise) product of two matrices A and B

Bibliography

Alter, O., Brown, P. O., and Botstein, D. (2000). Singular value decomposition for genome-wide expression data processing and modeling. *Proceedings of the National Academy of Sciences 97*(18), 10101–10106.

Avalos-Pacheco, A., Rossell, D., and Savage, R. S. (2018). Heterogeneous large datasets integration using Bayesian factor regression. *arXiv:1810.09894*, 1–46.

Avio, C. G., Gorbi, S., Milan, M., Benedetti, M., Fattorini, D., d'Errico, G., Pauletto, M., Bargelloni, L., and Regoli, F. (2015). Pollutants bioavailability and toxicological risk from microplastics to marine mussels. *Environmental Pollution 198*, 211 – 222.

Bar, H., Booth, J., and Wells, M. T. (2018). A scalable empirical Bayes approach to variable selection in generalized linear models. *arXiv:1803.09735*, 1–20.

Benito, M., Parker, J., Du, Q., Wu, J., Xiang, D., Perou, C. M., and Marron, J. S. (2004). Adjustment of systematic microarray data biases. *Bioinformatics 20*(1), 105–114.

Bentink, S., Haibe-Kains, B., Risch, T., Fan, J.-B., Hirsch, M. S., Holton, K., Rubio, R., April, C., Chen, J., Wickham-Garcia, E., Liu, J., Culhane, A., Drapkin, R., Quackenbush, J., and Matulonis, U. A. (2012, 02). Angiogenic mRNA and microRNA gene expression signature predicts a novel subtype of serous ovarian cancer. *PLOS ONE 7*(2), 1–9.

Berger, J. (1980, 07). A robust generalized bayes estimator and confidence region for a multivariate normal mean. *Ann. Statist. 8*(4), 716–761.

Bersanelli, M., Mosca, E., Remondini, D., Giampieri, E., Sala, C., Castellani, G., and Milanesi, L. (2016). Methods for the integration of multi-omics data: mathematical aspects. *BMC Bioinformatics 17*(2), 167–177.

Burges, C. J. C. (2010). Dimension reduction: A guided tour. *Foundations and Trends in Machine Learning 2*(4), 276–365.

Cancer Research UK (2015). Cancer Statistics for the UK. http://www.cancerresearchuk.org/health-professional/cancer-statistics. Accessed: September 2015.

Carvalho, C., Polson, N., and Scott, J. (2009). Handling sparsity via the horseshoe. *Journal of Machine Learning Research 5*, 73–80.

Carvalho, C. M., Chang, J., Lucas, J. E., Nevins, J. R., Wang, Q., and West, M. (2008). High-dimensional sparse factor modeling: Applications in gene expression genomics. *Journal of the American Statistical Association 103*(484), 1438–1456.

Chekouo, T., Stingo, F. C., Doecke, J. D., and Do, K.-A. (2015). mirna–target gene regulatory networks: A bayesian integrative approach to biomarker selection with application to kidney cancer. *Biometrics 71*(2), 428–438.

Choi, J.-H., Hong, S.-E., and Woo, H. G. (2017). Pan-cancer analysis of systematic batch effects on somatic sequence variations. *BMC Bioinformatics 18*(1), 211.

Collazo, R. A. and Smith, J. Q. (2016, 12). A new family of non-local priors for chain event graph model selection. *Bayesian Anal. 11*(4), 1165–1201.

Consonni, G. and La Rocca, L. (2011). *Moment Priors for Bayesian Model Choice with Applications to Directed Acyclic Graphs*, 119––14. Oxford: Oxford University Press.

Cox, D. R. (1972). Regression models and life-tables. *Journal of the Royal Statistical Society, Series B: Methodological 34*, 187–220.

Cunningham, J. P. and Ghahramani, Z. (2015). Linear dimensionality reduction: Survey, insights, and generalizations. *Journal of Machine Learning Research 16*, 2859–2900.

De Vito, R., Bellio, R., Trippa, L., and Parmigiani, G. (2018a). Bayesian multi-study factor analysis for high-throughput biological data. *arXiv:1806.09896*, 1–35.

De Vito, R., Bellio, R., Trippa, L., and Parmigiani, G. (2018b). Multi-study factor analysis. *arXiv:1611.06350*, 1–26.

Dempster, A. P., Laird, N. M., and Rubin, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Satistical Society, Series B: Statistical Methodology 39*(1), 1–38.

Dunson, D. and Bhattacharya, A. (2011). Sparse Bayesian infinite factor models. *Biometrika 98*, 291–306.

Earnshaw, R. (2017). *State of the Art in Digital Media and Applications*. Bookmetrix.

Fan, J. and Li, R. (2001). Variable selection via nonconcave penalized likelihood and its oracle properties. *Journal of the American Statistical Association 96*(456), 1348–1360.

Ferriss, J. S., Kim, Y., Duska, L., Birrer, M., Levine, D. A., Moskaluk, C., Theodorescu, D., and Lee, J. K. (2012, 02). Multi-gene expression predictors of single drug responses to adjuvant chemotherapy in ovarian carcinoma: Predicting platinum resistance. *PLOS ONE 7*(2), 1–9.

Fortin, J.-P., Sweeney, E. M., Muschelli, J., Crainiceanu, C. M., and Shinohara, R. T. (2016). Removing inter-subject technical variability in magnetic resonance imaging studies. *NeuroImage 132*, 198–212.

Fruchter, B. (1955). An introduction to factor analysis. *The American Journal of Psychology 68*, 1.

Frühwirth-Schnatter, S. and Lopes, H. F. (2018). Sparse Bayesian factor analysis when the number of factors is unknown. *arXiv:1804.04231*, 1–34.

Fúquene, J., Steel, M., and Rossell, D. (2018). On choosing mixture components via non-local priors. *arXiv:1604.00314*, 1–72.

Ganzfried, B. F., Riester, M., Haibe-Kains, B., Risch, T., Tyekucheva, S., Jazic, I., Wang, X. V., Ahmadifar, M., Birrer, M., Parmigiani, G., Huttenhower, C., and Waldron, L. (2013). curatedovariandata: Clinically annotated data for the ovarian cancer transcriptome. *Database 2013*.

George, E. and McCulloch, R. (1993). Variable selection via Gibbs sampling. *Journal of the American Statistical Association 88*(423), 881–889.

George, E. and McCulloch, R. (1997). Approaches for Bayesian variable selection. *Statistica Sinica*, 339–374.

Geweke, J. and Zhou, G. (1996). Measuring the Pricing Error of the Arbitrage Pricing Theory. CEMA Working Papers 276, China Economics and Management Academy, Central University of Finance and Economics.

Ghahramani, Z. and Beal, M. J. (2000). Variational inference for Bayesian mixtures of factor analysers. In S. A. Solla, T. K. Leen, and K. Müller (Eds.), *Advances in Neural Information Processing Systems 12*, 449–455. MIT Press.

Gligorijevic, V. and Przulj, N. (2015). Methods for biological data integration: perspectives and challenges. *Journal of the Royal Society Interface 12*.

Goh, W. W. B., Wang, W., and Wong, L. (2017). Why batch effects matter in omics data, and how to avoid them. *Trends in Biotechnology 35*, 498–507.

Griffiths, T. L. and Ghahramani, Z. (2011, July). The Indian Buffet Process: An introduction and review. *Journal of Machine Learning Research 12*, 1185–1224.

Hackstadt, A. J. and Hess, A. M. (2009). Filtering for increased power for microarray data analysis. *BMC Bioinformatics 10*(1), 11.

Harrell Jr., F. E., Califf, R. M., Pryor, D. B., Lee, K. L., and Rosati, R. A. (1982). Evaluating the yield of medical tests. *JAMA 247*(18), 2543–2546.

Hastie, T., Tibshirani, R., and Friedman, J. (2001). *The Elements of Statistical Learning*. Springer Series in Statistics. New York, NY, USA: Springer New York Inc.

Hirose, K. and Yamamoto, M. (2015). Sparse estimation via nonconcave penalized likelihood in factor analysis model. *Statistics and Computing 25*(5), 863–875.

Hirose, K., Yamamoto, M., and Nagata, H. (2016). *fanc: Penalized Likelihood Factor Analysis via Nonconvex Penalty*. R package version 2.2.

Hornung, R., Boulesteix, A.-L., and Causeur, D. (2016). Combining location-and-scale batch effect adjustment with data cleaning by latent factor adjustment. *BMC Bioinformatics 17*(1), 1–19.

Johnson, R. A. and Wichern, D. W. (Eds.) (1988). *Applied Multivariate Statistical Analysis*. Upper Saddle River, NJ, USA: Prentice-Hall, Inc.

Johnson, V. E. and Rossell, D. (2010). On the use of non-local prior densities in Bayesian hypothesis tests. *Journal of the Royal Statistical Society Series B: Statistical Methodology 72*(2), 143–170.

Johnson, V. E. and Rossell, D. (2012). Bayesian model selection in high-dimensional settings. *Journal of the American Statistical Association 107*(498), 649–660.

Johnson, W. E. and Li, C. (2009). *Adjusting Batch Effects in Microarray Experiments with Small Sample Size Using Empirical Bayes Methods*, 113–129. John Wiley & Sons, Ltd.

Johnson, W. E., Li, C., and Rabinovic, A. (2007). Adjusting batch effects in microarray expression data using empirical Bayes methods. *Biostatistics (Oxford, England) 8*(1), 118–27.

Kaiser, H. F. (1958, Sep). The varimax criterion for analytic rotation in factor analysis. *Psychometrika 23*(3), 187–200.

Kao, Y. and Roy, B. V. (2013). Learning a factor model via regularized PCA. *Machine Learning 91*(3), 279–303.

Kendall, M. (1975). *Multivariate Analysis*. London: Griffin.

Klami, A., Virtanen, S., Leppäaho, E., and Kaski, S. (2015). Group factor analysis. *IEEE Transactions on Neural Networks and Learning Systems 26*(9), 2136–2147.

Knowles, D. A. and Ghahramani, Z. (2011). Nonparametric Bayesian sparse factor models with application to gene expression modeling. *The Annals of Applied Statistics 5*(2B), 1534–1552.

Kristensen, V. N., Lingjaerde, O. C., Russnes, H. G., Vollan, H. K., Frigessi, A., and Borresen-Dale, A.-L. (2014, April). Principles and methods of integrative genomic analyses in cancer. *Nat Rev Cancer 14*(5), 299–313.

Lazar, C., Meganck, S., Taminau, J., Steenhoff, D., Coletta, A., Molter, C., Weiss-Solís, D. Y., Duque, R., Bersini, H., and Nowé, A. (2013). Batch effect removal methods for microarray gene expression data integration: a survey. *Briefings in Bioinformatics 14*(4), 469–490.

Leek, J. T., Johnson, W. E., Parker, H. S., Fertig, E. J., Jaffe, A. E., Storey, J. D., Zhang, Y., and Torres, L. C. (2017). *sva: Surrogate Variable Analysis*. R package version 3.26.0.

Leek, J. T., Scharpf, R. B., Bravo, H. C., Simcha, D., Langmead, B., Johnson, W. E., Geman, D., Baggerly, K., and Irizarry, R. A. (2010, October). Tackling the widespread and critical impact of batch effects in high-throughput data. *Nat Rev Genet 11*(10), 733–739.

Leek, J. T. and Storey, J. D. (2007, 09). Capturing heterogeneity in gene expression studies by surrogate variable analysis. *PLoS Genet 3*(9), 1–12.

Li, C. and Wong, W. H. (2001). Model-based analysis of oligonucleotide arrays: Expression index computation and outlier detection. *Proceedings of the National Academy of Sciences 98*(1), 31–36.

Li, C. and Wong, W. H. (2003). *DNA-Chip Analyzer (dChip)*, 120–141. New York, NY: Springer New York.

Lopes, H. F. and West, M. (2004). Bayesian model assessment in factor analysis. *Statistica Sinica 14*, 41–67.

Mitchell, T. J. and Beauchamp, J. J. (1988). Bayesian variable selection in linear regression. *Journal of the American Statistical Association 83*(404), 1023–1032.

NHS (2015). Overview, cancer. http://www.nhs.uk/conditions/cancer/Pages/Introduction.aspx. Accessed: September 2015.

Ni, Y., Stingo, F. C., and Baladandayuthapani, V. (2018). Bayesian graphical regression. *Journal of the American Statistical Association 0*(0), 1–14.

Novoradovskaya, N., Whitfield, M. L., Basehore, L. S., Novoradovsky, A., Pesich, R., Usary, J., Karaca, M., Wong, W. K., Aprelikova, O., Fero, M., Perou, C. M., Botstein, D., and Braman, J. (2004). Universal reference rna as a standard for microarray experiments. *BMC Genomics 5*(1), 20.

Olivetti, E., Greiner, S., and Greiner, S. (2012). ADHD diagnosis from multiple data sources with batch effects. *Frontiers in Systems Neuroscience 6*, 1662–5137.

Papaspiliopoulos, O. and Rossell, D. (2017). Bayesian block-diagonal variable selection and model averaging. *Biometrika 104*(2), 343–359.

Parker, H. S., Corrada Bravo, H., and Leek, J. T. (2014, September). Removing batch effects for prediction problems with frozen surrogate variable analysis. *PeerJ 2*, e561.

Pearson, K. (1901). Liii. on lines and planes of closest fit to systems of points in space. *Philosophical Magazine Series 6 2*(11), 559–572.

Perry, P. O. and Owen, A. B. (2010). A rotation test to verify latent structure. *Journal of Machine Learning Research 11*, 603–624.

Rhodes, D. R., Yu, J., Shanker, K., Deshpande, N., Varambally, R., Ghosh, D., Barrette, T., Pandey, A., and Chinnaiyan, A. M. (2004). Large-scale meta-analysis of cancer microarray data identifies common transcriptional profiles of neoplastic transformation and progression. *Proceedings of the National Academy of Sciences of the United States of America 101*(25), 9309–9314.

Rossell, D. and Rubio, F. J. (2018). Tractable bayesian variable selection: Beyond normality. *Journal of the American Statistical Association 0*(0), 1–17.

Rossell, D. and Telesca, D. (2017). Nonlocal priors for high-dimensional estimation. *Journal of the American Statistical Association 112*(517), 254–265.

Rossell, D., Telesca, D., and Johnson, V. (2013). High-dimensional bayesian classifiers using non-local priors. *Studies in Classification, Data Analysis, and Knowledge Organization*, 305–313.

Ročková, V. and George, E. I. (2014). EMVS: The EM approach to Bayesian variable selection. *Journal of the American Statistical Association 109*(506), 828–846.

Ročková, V. and George, E. I. (2017). Fast Bayesian factor analysis via automatic rotations to sparsity. *Journal of the American Statistical Association 111*(516), 1608–1622.

Ročková, V. and George, E. I. (2018). The Spike-and-Slab LASSO. *Journal of the American Statistical Association 113*(521), 431–444.

Schadt, E. E., Li, C., Ellis, B., and Wong, W. H. (2001). Feature extraction and normalization algorithms for high-density oligonucleotide gene expression array data. *Journal of Cellular Biochemistry 84*(S37), 120–125.

Scherer, A. (2009). *Batch Effects and Noise in Microarray Experiments: Sources and Solutions*. Wiley Series in Probability and Statistics. Wiley.

Schröeder, M. S., Culhane, A., Quackenbush, J., and Haibe-Kains, B. (2011). survcomp: an R/Bioconductor package for performance assessment and comparison of survival models. *Bioinformatics 27*(22), 3206–3208.

Schwarz, G. (1978, 03). Estimating the dimension of a model. *Ann. Statist. 6*(2), 461–464.

Seber, G. (1984). *Multivariate observations*. Wiley series in probability and mathematical statistics. New York, NY: Wiley.

Shah, M., Xiao, Y., Subbanna, N., Francis, S., Arnold, D. L., Collins, D. L., and Arbel, T. (2011). Evaluating intensity normalization on MRIs of human brain with multiple sclerosis. *Medical Image Analysis 15*(2), 267 – 282.

Shi, G., Lim, C. Y., and Maiti, T. (2018). Model selection using mass-nonlocal prior. *Statistics & Probability Letters*.

Shinohara, R. T., Sweeney, E. M., Goldsmith, J., Shiee, N., Mateen, F. J., Calabresi, P. A., Jarso, S., Pham, D. L., Reich, D. S., and Crainiceanu, C. M. (2014). Statistical normalization techniques for magnetic resonance imaging. *NeuroImage : Clinical 6*, 9–19.

Sims, A. H., Smethurst, G. J., Hey, Y., Okoniewski, M. J., Pepper, S. D., Howell, A., Miller, C. J., and Clarke, R. B. (2008, Sep). The removal of multiplicative, systematic bias allows

integration of breast cancer gene expression datasets – improving meta-analysis and prediction of prognosis. *BMC Medical Genomics 1*(1), 42.

Spearman, C. (1904). "general intelligence," objectively determined and measured. *The American Journal of Psychology 15*(2), 201–292.

Strawderman, W. E. (1971, 02). Proper Bayes minimax estimators of the multivariate Normal mean. *Ann. Math. Statist. 42*(1), 385–388.

Therneau, T. M. (2015). *A Package for Survival Analysis in S*. version 2.38.

Tibshirani, R. (1994). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society, Series B 58*, 267–288.

Tipping, M. E. and Bishop, C. M. (1999). Probabilistic principal component analysis. *Journal of the Royal Statistical Society, Series B 61*, 611–622.

Ulfarsson, M. O. and Solo, V. (2008). Sparse Variable PCA Using Geodesic Steepest Descent. *IEEE Transactions on Signal Processing 56*(12).

Wan, Y.-W., Allen, G. I., Anderson, M. L., and Liu, Z. (2015). *TCGA2STAT: Simple TCGA Data Access for Integrated Statistical Analysis in R*. R package version 1.2.

Wan, Y.-W., Allen, G. I., and Liu, Z. (2016). TCGA2STAT: simple TCGA data access for integrated statistical analysis in R. *Bioinformatics 32*(6), 952–954.

Wang, J. and Zhao, Q. (2015). *cate: High Dimensional Factor Analysis and Confounder Adjusted Testing and Estimation*. R package version 1.0.4.

West, M. (2003). Bayesian factor regression models in the "large p, small n" paradigm. In *Bayesian Statistics 7*, 723–732. Oxford University Press.

Witten, D. M., Tibshirani, R., and Hastie, T. (2009). A penalized matrix decomposition, with applications to sparse principal components and canonical correlation analysis. *Biostatistics 10*(3), 515–534.

World Health Organization (2014). World cancer report 2014. Technical report.

Yang, Y. H., Dudoit, S., Luu, P., Lin, D. M., Peng, V., Ngai, J., and Speed, T. P. (2002). Normalization for cDNA microarray data: a robust composite method addressing single and multiple slide systematic variation. *Nucleic Acids Research 30*(4), e15.

Zhang, C.-H. (2010, 04). Nearly unbiased variable selection under minimax concave penalty. *Ann. Statist. 38*(2), 894–942.

Zhu, B., Song, N., Shen, R., Arora, A., Machiela, M. J., Song, L., Landi, M. T., Ghosh, D., Chatterjee, N., Baladandayuthapani, V., and Zhao, H. (2017). Integrating clinical and multiple omics data for prognostic assessment across human cancers. *Scientific Reports 7*(1), 16954.